Masters Theses 1911 - February 2014

2009

Dual-Process Theory and Syllogistic Reasoning: A Signal Detection Analysis

Chad M. Dube
University of Massachusetts Amherst

DUAL-PROCESS THEORY AND SYLLOGISTIC REASONING: A SIGNAL DETECTION ANALYSIS

A Thesis Presented

by

Chad M. Dube

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

February 2009

Psychology

DUAL-PROCESS THEORY AND SYLLOGISTIC REASONING: A SIGNAL DETECTION ANALYSIS

A Thesis Presented

by

Chad M. Dube

Approved as to style and content by:

Caren M. Rotello, Chair

Neil A. Macmillan, Member

Marvin W. Daehler, Member

Melinda A. Novak, Department Head
Department of Psychology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Overview

Galotti (1989) defines reasoning as "…mental activity that consists of transforming given information (called the set of premises) in order to reach conclusions." Though the focus of the research to be described herein is not to debate human rationality, that debate (see, e.g., Shafir & LeBoeuf, 2002; Stanovich & West, 2000) has highlighted the difficulty of adequately defining reasoning. In particular, Stanovich and West's (2000) review of the rationality debate includes commentary from the standpoint of evolutionary psychology that suggests subjects' systematically poor performance on logical tasks is often consistent with what would be the most utile response in the everyday world. The evolutionary suggestion raises a question as to whether 'reasoning' is best thought of as what logicians do or as what most people do in their day-to-day lives. Correct responses to reasoning problems, both in this review and the research to be reported, are the ones expected by normative theorists, i.e., by the logician, though whether subjects are behaving rationally when they do so (or fail to do so) is of no concern. For the sake of simplicity then, I will assume human reasoning is as Galotti (1989) describes it.

Traditionally, logic distinguishes between two types of arguments: inductive and deductive (Copi & Cohen, 1994). Inductive arguments, generally speaking, involve making generalizations given a relatively limited set of information. The following is an example of a valid categorical induction problem; the solution of this problem requires the subject to reason probabilistically by combining the information in the premises with everyday knowledge.

All cows are mammals and have lungs.

All whales are mammals and have lungs.

All humans are mammals and have lungs.

Probably all mammals have lungs. (1)

Deductive arguments are distinguished from inductive ones in that the only deductively valid conclusions are those that do not invoke information beyond that which is contained in the premises. Conditional reasoning is an example of deductive logic. Conditional problems generally state a rule of the form 'if p then q', followed by a truth statement about either p or q. The reasoner must indicate whether a conclusion can be drawn linking p and q. The important point is that in this case the conclusion is only valid if it is necessitated by the premises. The following is an example of a valid conditional reasoning problem.

If Socrates is human then Socrates is mortal.

Socrates is human.

Socrates is mortal. (2)

The distinction between induction and deduction has also been adopted by psychologists (Evans, 2007; Heit, 2007). A conservative view is that this distinction applies only to the stimuli themselves, and that the same basic reasoning capacity or mechanism is invoked when a subject attempts to solve inductive and deductive problems. A more radical view is that induction and deduction also map onto qualitatively different underlying processes. The view that there are two reasoning systems has had a considerable impact on the reasoning literature (Evans, 2007; Sloman, 1996; Stanovich & West, 2000).

The focus of this research is to evaluate claims that two reasoning systems contribute to subjects' responses in reasoning experiments involving syllogisms, which are a type of deductive argument used widely in research related to this question. It is also is important to know whether similar conclusions that have been reached in experiments employing inductive stimuli, such as categorical induction problems, generalize to experiments that use deductive stimuli such as syllogisms. Inferential and descriptive techniques developed within the well-established signal detection framework (Green & Swets, 1966; Macmillan & Creelman, 2005) will be applied to data collected from two syllogistic reasoning experiments, extending previous work by Heit and Rotello (2005), to be described below.

Syllogistic Reasoning

A great deal of research in the area of deductive reasoning has used syllogisms as stimuli. Syllogisms are logical arguments consisting of two premises and a conclusion, which may or may not follow logically from the premises. The task of the subject is to deduce a conclusion by linking the Z and X terms, referred to as subject and predicate, by way of their relationships to the middle term. An example of a syllogism is the following (valid) argument, adapted from Johnson-Laird and Steedman (1978):

All artists are beekeepers

No beekeepers are chemists

No chemists are artists (3)

Syllogisms may contain concrete or abstract content. An abstract version of (3) might be the following:

All X are Y

No Y are Z

No Z are X (4)

Three versions of the syllogistic reasoning task are commonly used: conclusion evaluation, forced-choice, and conclusion production. Subjects in a conclusion evaluation experiment typically receive examples like (3) and are asked whether the conclusion they are given follows necessarily from the premises. Subjects in the forced-choice experiment must choose a conclusion from a set of possibilities that includes ‘no valid conclusion.’ Subjects in a production task typically receive a set of premises, and are asked to either respond with a conclusion of their own or to indicate that no valid conclusion can be drawn.

The building blocks of syllogisms have been shown to affect the number and the nature of errors subjects commit in attempting to solve them (Dickstein, 1978; Johnson-Laird, 1983). One such factor is quantification. Traditionally, each sentence of the syllogism can take one of four quantifiers: 'All,' 'No,' 'Some,' and 'Some...are not,' labeled A, E, I, and O, respectively. An early finding in the literature was that certain combinations of premise quantifiers can bias the subject in favor of particular quantifiers in the conclusion; this is known as the atmosphere effect (Woodworth & Sells, 1935; Sells, 1936). Begg and Denny (1969) summed up atmosphere biases with two predictive heuristics:

1. If there is at least one negative premise ('No' or 'Some...are not'), favor a negative conclusion; otherwise, favor a positive conclusion ('All' or 'Some').
2. If there is at least one particular premise ('Some' or 'Some...are not'), favor a particular conclusion; otherwise, favor a universal conclusion ('All' or 'No').

A second effect of quantification is illicit conversion (Dickstein, 1975; 1981; Revlis, 1975). For example, Revlis (1975) pointed out that subjects confronted with relations such as 'All A are B' may erroneously infer 'All B are A' to be true as well, and that on some syllogisms in which invalid conclusions are drawn the response may be perfectly valid if one assumes the converted version of the premise(s) in question. Subsequent research demonstrated that error rates can be substantially reduced when instruction is given in logical interpretation of nonconvertible quantifiers (Dickstein, 1975).

Another important factor in the difficulty of syllogisms is figure, which is the combined ordering of terms in the first and second premises. Since there are two terms per premise, the arrangement yields four possible syllogistic figures, illustrated in Figure 1.

Holding the order of conclusion terms constant (i.e., X-Z or Z-X), there are 4 possible quantifiers per premise, and 4 possible figures, which yields 4 x 4 x 4 = 64 possible syllogisms. As pointed out by Johnson-Laird (1983), allowing the ordering of conclusion terms to vary yields a much larger set of 256 possible syllogisms.

A landmark experiment by Dickstein (1978), using a five-alternative forced-choice paradigm and Z-X conclusions, demonstrated that many erroneous responses in syllogistic reasoning could be accounted for by the relationship between the ordering of terms in the premises and that of the terms in the conclusion. More specifically, accuracy for valid syllogisms in figure 1 was higher than for valid syllogisms in figure 4, with 2 and 3 intermediate between the two. Dickstein argued this was because a valid Z-X conclusion is consistent with the ordering of premise terms in figure 1, while in figure 4 it is in the opposite direction, which requires 'backward processing' on the part of the subject and imposes a greater strain on working memory.

When the figure or quantification of a syllogism contributes to the difficulty of its solution, the effect is referred to as *structural*. Another source of difficulty is the content of the problem. Content effects arise when concrete problems are used, and the quantifiers invoke relations between terms that may or may not arise in the real world. An example of a pervasive content effect is belief bias (e.g. Cherubini, Garnham, & Morley, 1998; Evans, Newstead, & Byrne, 1993; Evans, Handley, & Harper, 2001; Markovits & Nantel, 1989; Roberts & Sykes, 2003; Shynkaruk & Thompson, 2006), which is a tendency on the part of the subject to reject or accept potential conclusions on the basis of consistency with prior beliefs, regardless of logical status. Consider, for example, the following problem (cf. Evans, Barston, & Pollard, 1983):

No addictive things are inexpensive.

Some cigarettes are inexpensive.

Some cigarettes are not addictive. (5)

This syllogism is logically valid, but its conclusion is unbelievable. An example of the converse, an invalid believable problem, would be as follows:

No addictive things are inexpensive.

Some cigarettes are inexpensive.

----

*Some addictive things are not cigarettes. (6)

Belief bias effects are notoriously difficult to overcome, with even the most meticulous and extensive logical instruction only serving to reduce, but not eliminate, the effect (Evans, Newstead, Allen, & Pollard, 1994).

Evans, Barston, and Pollard (1983) conducted an investigation into the belief bias effect which was notable in that it ruled out the known structural factors (Revlis, 1975; Revlin, Leirer, Yopp, & Yopp, 1980). Subjects were presented with four types of arguments in which the validity and believability of the conclusion were crossed; they were asked to judge whether the conclusion was valid. Conversion was controlled by only using the logically convertible quantifiers 'Some' and 'No', and atmosphere was controlled by only using 'Some...are not' conclusions, which are favored by the bias. In two of the three experiments, figure was controlled for by using both Z-X and X-Z conclusions for each problem. In these experiments, only figures 2 and 3 were used, for which Dickstein (1978) found no clear preference in terms of conclusion direction. The design and results are summarized in Table 1.

Evans et al. (1983) obtained three effects which have since been replicated in a number of studies. Subjects accepted more valid than invalid conclusions, and more believable than unbelievable conclusions. Most importantly, there was an interaction between logic and belief, such that the difference in acceptance of believable and unbelievable problems was greater when problems were invalid than when they were valid. The effect appears to stem from the very low acceptance rate of invalid unbelievable problems, though the precise nature of the Evans et al. result is unclear. In particular, it is not clear whether the effect is primarily due to logical processing pre-empted by belief status, belief-based responding pre-empted by logical status, or some mixture of the two. As will soon be clear, explaining the interaction has been a major goal of extant theories of belief bias.

Theories of Belief Bias

Selective Scrutiny

Several explanations of the findings in Evans et al. (1983) have been proposed. The first of these was originally suggested by the authors themselves, and was subsequently termed the selective scrutiny model. Selective scrutiny predicts that subjects focus initially on the conclusion of the argument, and accept believable conclusions without considering the logic of the argument. When conclusions are not believable, subjects then reason through the premises and accept or reject conclusions on the basis of their perceived logical validity. Selective scrutiny could thus be seen as a process whereby logic-based responding is driven by the believability of conclusions (belief→logic); the belief x logic interaction is accounted for in that reasoning only occurs when syllogisms are unbelievable. While more recent work does appear to support the idea that conclusion believability has an influence on the processing of premises (e.g., Ball, Philips, Wade, & Quayle, 2006; Morley, Evans, & Handley, 2004), the theory by itself cannot account for main effects of logic on believable problems (see Klauer, Musch, & Naumer, 2000 for a meta-analysis).

Misinterpreted Necessity

A second theory proposed by Evans et al. (1983) that has since gained substantial attention in the literature is the misinterpreted necessity model (Markovits & Nantel, 1989; Newstead, Pollard, Evans, & Allen, 1992). Misinterpreted necessity predicts, in contrast to selective scrutiny, that subjects will engage in reasoning at the outset, and only rely on belief after reaching conclusions that are consistent with, but not necessitated by, the premises. An example of this state of affairs is given by the following problem (7):

Some X are Y

No Z are Y

----

*Some Z are not X (7)

Specifically, subjects are said to misunderstand the notion of necessity, and to become confused or uncertain when they are confronted with conclusions that they know to be consistent with but not necessitated by the premises. Misinterpreted necessity, one might argue, views belief-based responding as an escape-hatch mechanism (logic→belief), and provides a sensible explanation of the finding of increased sensitivity to belief on invalid problems since the only problems that can lead to indeterminate conclusions are by definition invalid ones.

Newstead et al. (1992) provided evidence both for and against misinterpreted necessity. Across two initial experiments, they varied whether conclusions were determinately or indeterminately invalid and only obtained the interaction when problems were of the latter variety. In a third experiment, however, the logic x belief interaction was not obtained despite the use of indeterminately invalid problems. The reason for this apparent inconsistency will become clear shortly. A further weakness of the misinterpreted necessity model is its inability to account for effects of belief on valid problems (Klauer et al., 2000; Newstead et al., 1992).

Mental Models

A third theory of belief bias follows from the mental models framework originally proposed by Johnson-Laird and colleagues (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978). Mental models theories of belief bias (Oakhill & Johnson-Laird, 1985; Oakhill, Johnson-Laird, & Garnham, 1989) generally assume three basic stages in the processing of syllogisms. First, subjects construct a mental representation that integrates the premises, the terms of which are described more or less as mental tokens. Second, subjects check to see whether the conclusion is consistent with the model they have constructed. If the conclusion is not consistent, it is rejected; if the conclusion is consistent, the subject evaluates its believability. If a conclusion is believable, it is accepted; if a conclusion is unbelievable, a third process is initiated the goal of which is to construct alternative models of the premises. If the conclusion is consistent with all alternative models, it is accepted; if the conclusion is not consistent with all models, it is rejected. Mental models theory essentially proposes that responses result from a mixture of belief- and logic-based operations, rather than a single linear relation. An illustration of this process is provided in Figure 2.

The mental models explanation can account for the fact that subjects are more sensitive to belief on invalid problems. The theory classes problems according to the number of possible models of the premises they allow; there are single- and multiple-model problems. Specifically, the role of believability is that it biases the reasoning process itself, such that construction of alternative models only occurs for unbelievable problems, and this manifests itself as a greater effect of logic when problems are unbelievable. A clear prediction of mental models is that the belief x logic interaction will only occur for stimuli that allow the generation of alternative models (multiple-model problems), irrespective of the determinacy status of the conclusion. This is the manipulation carried out by Newstead et al. (1992) in experiment 3, mentioned above: the stimuli were single-model, indeterminately invalid problems, and no interaction was obtained, consistent with the mental models interpretation.

While mental models theory is compelling, it is important to note that it was originally developed to explain conclusion production data, and as such it has been argued by some researchers that it may not accurately characterize the evaluation paradigm, which seems to require different processes and to inspire different biases. For instance, Morley et al. (2004) evaluated the hypothesis that conclusion production encourages 'forward' reasoning (from premises to conclusion) while conclusion evaluation encourages 'backward' reasoning (the conclusion biases construal of the premises). In a series of four experiments, Morley et al. demonstrated figural bias in the absence of belief bias in a conclusion production task, while the opposite (belief bias in the absence of figural bias) held for the conclusion evaluation task, consistent with their claims. The authors suggested that a mental models account in which models of premises are constructed can still apply, but that it would need to be modified to allow for effects of conclusions on the construction of those models.

Mental models theory also suffers from the fact that the belief x logic interaction has been obtained using one-model problems (Gilinsky & Judd, 1994; Klauer et al., 2000; Oakhill et al., 1989). Oakhill et al. (1989) responded to this issue by affixing an *ad hoc* conclusion filtering mechanism to their version of the mental models framework. In other words, subjects may be processing syllogisms the way mental models predicts, but in cases where conclusions are unbelievable subjects may still exhibit response biases that operate secondarily to filter (reject) such conclusions. Even if one were to maintain the conclusion filter, more recent findings from eyetracking (Ball et al., 2006) and response time (Thompson et al., 2003) experiments have converged on the notion that subjects actually spend more time processing believable and valid problems than unbelievable and invalid ones, which is inconsistent with the alternative generation account of the interaction. Though it could be argued that the above measures are contaminated by wrap up effects (e.g. Hirotani, Frazier, & Rayner, 2006), it is clear that the data so far do not clearly favor the mental models interpretation.

Overall, it appears that though each of the theories may account for some of the data, none of them provides a systematic account of all findings related to the belief bias effect. A more general account may be found in dual-process theory, the third conception alluded to by Evans et al. (1983).

Dual-Process Theory

Stanovich and West (2000) summarized and illustrated the influence of dual-process theories, which have gained widespread attention in the reasoning literature (e.g. Beller & Spada, 2003; Chater & Oaksford, 2001; Evans, 2003, 2007; Feeney, 2007; Markovits & Schroyens, 2007; Shafir & LeBoeuf; 2002; Sloman, 1996). The authors discussed a number of findings from a wide array of reasoning paradigms, and provided a meta-theoretical summary of the conclusions reached by researchers in those areas. Many of the conclusions are similar to one another in that they specify two mechanisms, the characteristics of which appear to fall into distinct categories (see Table 2). Stanovich and West referred to these categories as system 1 and system 2.

System 1 processes are characterized as fast-acting, heuristic-based, associative processes. They are the 'quick and dirty' processes that often produce errors such as the acceptance of fallacies in logical arguments. System 2 processes, on the other hand, are slower, more analytic processes, and are thought to require decontextualized processing which ignores or inhibits knowledge-based biases. Though the generality of Stanovich and West's categorical distinction and the inclusion of the various theories subsumed by it may be questioned, it is possible that a general framework such as this may apply to more specific problems in the reasoning literature.

Evans and Curtis-Holmes (2005) evaluated dual-process theory as a potential explanation of the belief bias effect. Specifically, the authors hypothesized system 1 processes to be driving belief-based responding, while system 2 processing was theorized to drive logic-based responding. Belief bias, according to dual-process theory, is an example of a conflict between these two systems of responding, and this is reflected in the data as effects of belief and logic. A desirable state of affairs, then, is to create a set of conditions that could potentially distinguish between the two systems. One possibility is to constrain the operation of one system without necessarily hindering the other, e.g. by asking subjects to make speeded decisions. This was the manipulation carried out by Evans and Curtis-Holmes.

Subjects were divided into two groups: a deadline group and an unspeeded group. The deadline group was given up to 10 seconds to respond to syllogisms of the sort used by Evans et al. (1983). The authors argued that a 10 second deadline would be short enough to effectively reduce analytical processing, citing a finding from Thompson, Striemer, Reikoff, Gunter, and Campbell (2003) that subjects average over 20 seconds to evaluate similar problems. The second group was allowed unlimited time to evaluate the same problems. Results are reproduced in Figure 3. The standard effects were obtained in the unspeeded group, in line with the prediction of dual process theory that both systems ought to contribute in the usual fashion. In the deadline group, however, there were notable deviations from the usual findings. First, subjects were equally sensitive to belief on valid and invalid problems, in line with the hypothesis that a logic-based process was blocked by the deadline. Second, the deadline group was more sensitive to belief than was the unspeeded group, indicating greater reliance on system 1. Finally, subjects were less likely to discriminate between valid and invalid arguments in the deadline group, in line again with the initial prediction. Evans and Curtis-Holmes concluded that belief bias reflects the operation of two distinct systems of reasoning.

Neuroimaging data in favor of dual-process theory have also been obtained. Goel and Dolan (2003) used an event-related fMRI procedure to scan subjects while they evaluated syllogisms similar to those used by Evans and Curtis-Holmes (2005). The imaging data were analyzed in terms of four trial types: belief-neutral (all responses to problems with neutral content), belief-laden (all responses to believable and unbelievable problems), correct inhibitory (correct responses to valid unbelievable and invalid believable problems), and incorrect inhibitory (incorrect responses to valid unbelievable and invalid believable problems). Results are illustrated in Figure 4. Goel and Dolan found that trials in which logic and belief conflicted appeared to recruit executive control processes, in that regions of the prefrontal cortex associated with inhibitory control were activated, while those trials that did not entail conflict (belief-neutral trials) appeared to rely primarily on regions of the parietal lobe. The authors concluded that two distinct, dissociable systems appear to underlie responding in the belief bias task, consistent with the predictions of dual-process theory.

Inductive Reasoning and Dual-Process Theory

The belief bias task, often studied in deductive reasoning paradigms such as propositional (e.g. Markovits & Schroyens, 2007) and syllogistic reasoning, has also been used to argue for fundamentally different inductive and deductive systems, both operating on the processing of inductive stimuli (Rips, 2001). Rips' stimuli were conditional and categorical induction problems that varied in inductive strength (believability) and deductive correctness (validity). An example of a conflict problem similar to a syllogistic invalid believable problem is an argument like the following:

Grizzlies hibernate during January.

----

*Black bears hibernate during January. (8)

An example of a facilitatory, valid believable induction problem is:

Grizzlies hibernate during January, and black bears hibernate during January.

Grizzlies hibernate during January. (9)

As described by Rips (2001), a unitary view of the reasoning process indicates a single dimension of argument strength underlies the decisions subjects make; effects of strength and correctness simply reflect a shift in the criterion subjects use to judge the acceptability of arguments. For example, arguments judged by subjects to be valid or deductively correct are those arguments whose strength surpasses a relatively high criterion on the strength axis (see Figure 5). Arguments judged to be inductively strong only require enough strength to pass a lower criterion. In other words, the unitary view makes a prediction about the ordering of problems on the strength dimension in Figure 5: A>B>C.

Rips (2001) attempted to modulate inductive and deductive responding by manipulating instructions. One group of subjects received induction instructions which stressed the plausibility of arguments, and asked that the reasoner evaluate the strength of the arguments. A second group received deduction instructions which stressed the concept of logical necessity and asked the reasoner to evaluate the validity of the arguments (see Appendix A for actual instructions). All subjects were then presented with categorical induction problems in which levels of the strength and correctness factors were crossed. Figure 6A illustrates the results of Rips' experiment, a 3-way interaction between logic, belief, and instructions. Considering the results for the deduction group, Rips obtained a belief x logic interaction similar to the one found in studies of belief bias, in that belief had a greater effect on incorrect arguments. Though Rips did not directly compare the size of the interaction for induction and deduction, it appears to be larger for the induction than for the deduction group. The difference in effect size is due to a significant crossover effect on conflict problems: the deduction group gave more positive responses to correct and inconsistent problems than for incorrect and consistent ones (it depended primarily on deductive correctness), while the opposite pattern emerged in the induction group (it depended primarily on inductive strength). This finding, i.e., an inconsistent relationship between the groups on the same problems, is contrary to the necessary prediction of the unitary view that problems be ordered the same way on the strength dimension for both groups (A>B>C, Figure 5).

It is important to note that had the data not conformed to the ordering predicted by Rips' (2001) unitary model, a unitary view might still have accounted for them so long as that ordering was the same for the induction and deduction groups. The fact that the relationship between the groups changes sign as a function of problem type (Figure 6B) means the data fail to satisfy a necessary prediction of any single-process account: the relationship between two groups that respond on the basis of the same underlying process should be the same across all levels of a given predictor variable (Bamber, 1979). In other words, the function relating induction and deduction should be monotonic if a unitary view is correct. Rips rejected the unitary view and concluded inductive and deductive responses reflect distinct systems of reasoning. The nonmonotonic relationship reported by Rips lends weight to his conclusion as these analyses have been shown to effectively distinguish single- from multiple-process accounts even in situations in which other inferences based on functional dissociations can be misleading (Dunn & Kirsner, 1988).

Several questions arise if one accepts the view that two processes contribute to human reasoning in the research reviewed above. One class of questions regards the nature of the two systems. Is system 2 reasoning a continuous or an all-or-none process? How does it differ from system 1 reasoning? Answering these questions may also provide new information regarding the belief x logic interaction. For example, the theories mentioned above highlight the question of whether the greater effect of belief for invalid problems is actually due to logic-inspired belief-based responding, belief-inspired logic-based responding, or some mixture. New information pertaining to the nature of system 2 processing could be helpful in revealing whether there are particular patterns of system-based responding. Fortunately, there exists a powerful framework for dealing with this class of questions, one which has been largely neglected in the area of reasoning. It is desirable, if one is to accept a dual-process account of human reasoning, to obtain converging evidence by way of such a model.

Signal Detection Theory and ROC Analysis

In the area of recognition memory, a debate regarding whether a single- or dual-process account provides the best description of subjects' behavior has been ongoing for the past 30 years. Though the areas of memory and reasoning may be sufficiently distinct from one another to warrant caution in making comparisons, the goal of this research is not to generalize across these areas in terms of processes or specific theories. Rather, the goal is to describe an inferential and descriptive model that has been shown to provide important insights into the question of single- versus multiple-process accounts of recognition, with the aim of extending its application to the area of human reasoning.

Briefly, the standard item recognition paradigm involves the presentation of a list of words, followed by a test in which the subject must distinguish between previously studied words and new words, or lures. The recognition experiment yields four types of responses. If a test word is actually an old (previously studied) word, the subject's response is either a 'hit' (an 'old' response) or a 'miss' (a 'new' response). If a test word is actually a new word (lure), the subject's response is either a 'correct rejection' (a 'new' response) or a 'false alarm' (an 'old' response). Much of the research using this and related tasks has been guided by the use of signal detection theory, a theoretical and inferential framework that began to impact memory theorists in the 1960s (see Banks, 1970 for review), and continues to have a profound influence on models and theories of recognition to the present day (Kelly & Wixted, 2001; Rotello, Macmillan, & Reeder, 2004; Wixted, 2007; Yonelinas, 1994).

In its most basic form, detection theory[1] posits that memory decisions reflect the operation of a single, continuous 'memory strength' variable (see Figure 7). In the memory experiment described above, the memory strength of old and new items is

---

[1] Though detection theory may be extended to incorporate the operation of multiple continuous processes (Kelly & Wixted, 2001; Rotello, Macmillan, & Reeder, 2004), 'signal detection theory' in this writing will be used to refer solely to the more basic, univariate model.

distributed normally, and the ability to distinguish between them reflects heightened activation of old items (higher mean strength), as a result of recent study. The distance between the distribution means provides an index of sensitivity, which can be calculated using the *d'* parameter. *d'*, assuming the assumptions of normality and homogeneity of variance are met, is the difference between the *z*-transformed hit and false alarm rates of a given subject or group of subjects, and is independent of response bias.

$$d' = z(H) - z(F)$$

Response bias (willingness to say 'old') can be measured in a number of ways (see Macmillan & Creelman, 2005 for discussion), but the more common methods are all related by the criterion placement parameter. Criterion placement, *c*, reflects bias relative to the zero-bias point where the old and new item distributions cross over; liberal biases (maximizing hits at the cost of increasing false alarms) reflect negative values of *c*, while conservative biases (minimizing false alarms at the cost of a reduced hit rate) reflect positive values of *c*.

$$c = -.5(z(H) + z(F))$$

As illustrated in Figure 7, area under the old item distribution to the right of the criterion corresponds to the hit rate (*H*), while the area under the new item distribution to the right of the criterion corresponds to the false alarm rate (*F*). The area of overlap between the distributions reflects low sensitivity; the greater this area is relative to either distribution, the lower overall sensitivity will become, regardless of criterion placement. The areas under the old and new item distributions to the left of the criterion correspond to misses (*M*) and correct rejections (*CR*), respectively.

A powerful method for the evaluation of detection theory and other models, as well as for checking the assumptions of a given model, is the analysis of receiver-operating characteristics, or ROCs. The ROC plots hit rate as a function of false alarm rate at different levels of response bias. One very common method for collecting empirical ROC data is to require subjects to follow their responses (e.g. 'old' or 'new') with an indication of their confidence in the response on a rating scale. The ROC in Figure 8 below was plotted using a 6-point confidence scale, in which a 1 corresponded to 'sure old' and a 6 corresponded to 'sure new.' As a rating of 1 corresponds to the most stringent criterion for an 'old' response, both the hit and false alarm rate should at this point be lower than at any other point on the function. An important property of ROCs is that they are cumulative, i.e., the (*F*, *H*) pair at 2 is the sum of hit and false alarm proportions from confidence levels 1 and 2, the (*F*, *H*) pair at 3 is the sum of the proportions from 1 to 3, and so forth. The cumulative nature of the 6-point ROC results in a function with 5 points and an upper-x intercept at (1, 1).

The signal detection model, which assumes that normal, Gaussian distributions of strength underlie rate of responding, can be used to generate theoretical ROCs for a given level of sensitivity (isosensitivity curves). A 6-point ROC generated from such a model yields a curvilinear ROC that is symmetrical about the minor diagonal (see figure). Plotting the same ROC on *z*-coordinates reveals a linear function with a slope of 1, and the difference between $z(H)$ and $z(F)$ at the most stringent point on the *z*ROC will be equivalent to d' itself. In this way, sensitivity in the signal detection model is reflected by the height of the ROC in x, y space; the distance between the ROC and the major diagonal (which measures chance performance) increases as sensitivity increases.

Response bias is reflected in the points on the ROC, which correspond to different criteria on the strength axis; as one moves from a rating of 1 to 6, the criterion becomes increasingly liberal (moves farther to left), increasing both H and F. In this way, points on the same ROC reflect equal sensitivity but different levels of response bias. The theoretical ROC implied by signal detection theory, with its distinctive curvilinearity, was shown by researchers in the early decades of the tradition to provide a better fit to empirical ROCs than did other model-implied ROCs, such as those implied by threshold theory (e.g. Egan, 1958; Green & Swets, 1966).

The slope of the *z*ROC, which is equal to the ratio of new and old item standard deviations (σ_n/σ_o), can be used to make inferences about the variances of strength distributions (Figure 9). Assuming, e.g., σ_n is static, the slope of the ROC will decrease (or increase) as σ_o increases (or decreases). In memory experiments, *z*ROC slope is often less than one (Glanzer, Kim, Halford, & Adams, 1999; Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Ratcliff, McKoon, & Tindall, 1994). A series of item recognition experiments by Ratcliff et al. (1994), for instance, varied rate of presentation, list length, word frequency, presentation duration, and semantic similarity and found that in almost every instance *z*ROC slope remained constant at about .80. More recent experiments by Glanzer et al. (1999) and Heathcote (2003) also varied depth of encoding, number of repetitions, semantic concreteness, categorical relatedness, orthographic similarity, and category length (number of related words). The results from many of these experiments indicated that as recognition accuracy increases, slope decreases.

Effects on ROC indices such as slope or height across experimental conditions are consistent with multiple-process models like the one suggested by Rips (2001). In fact, in the memory literature slope effects were argued by Yonelinas (1994) to reflect the contribution of two qualitatively different memory processes to distributions of memory strength. In Yonelinas' (1994) dual-process framework, recollection, i.e., the retrieval of specific details related to the memory probe, is modeled as an all-or-none threshold component that is highly accurate and should only contribute to high-confidence memory judgments. At test, old items either pass a threshold and are recollected or they fail to do so and a second, strength-based signal detection process is used to output a decision. Strength-based decreases in slope, then, are said to reflect a growing subset of items whose strength has been boosted past the recollection threshold. This subset would produce a right-skewed old item distribution, decreasing σ_n/σ_o. Though the dual-process model remains very controversial (see Wixted, 2007, for review), it nonetheless serves to illustrate the importance of ROC indices in providing a window onto underlying processes.

Similar inferences can be made by applying signal detection and ROC analysis to reasoning data. In this case, parameters that memory theorists use to describe the strength of old and new items are used to describe the strength of valid and invalid arguments. The slope of the *z*ROC, then, reflects the ratio of σ_{invalid} to σ_{valid}. It is desirable to know whether slope will change in response to manipulations directed at system-based responding, as such differences could indicate a qualitative change in the form of the argument strength distributions. If, for instance, greater effects of belief on invalid problems result from a unique mixture of logic- and belief-based responding acting on invalid unbelievable arguments, this mixture might be expected to selectively affect the variance of invalid unbelievable arguments. Such effects would be reflected in a change in slope relative to *z*ROC slope for neutral or believable problems.

An additional concern that applies regardless of the particular paradigm one works with is whether the assumptions of a given model have been met. Specifically, without recourse to ROCs one may adopt equal-variance parameters when the data do not support that model's assumptions; this can greatly elevate the risk of committing a type I error (Rotello, Masson, & Verde, 2008). In the (frequently occurring) event that ROCs indicate the equal-variance assumption has been violated, an unequal-variance signal detection framework can be adopted. In this case the measures d_a and c_a may be substituted for *d'* and *c*, respectively. The unequal variance parameters are obtained by weighting *d'* and *c* by *s*, the standard deviation of the lure distribution (for derivation, see Macmillan & Creelman, 2005).

$$da = [2/(1 + s^2)]^{1/2} [z(H) - sz(F)]$$

$$c_a = [(-\sqrt{2})s]/[(1 + s^2)^{1/2}(1 + s)] [z(H) + z(F)]$$

A synthesis: Heit and Rotello (2005)

Heit and Rotello (2005) reported two experiments conducted with the aim of further evaluating Rips' (2001) dual-process conception of inductive reasoning using ROC methodology. In experiment 1, Heit and Rotello replicated Rips' experiment: subjects were given either induction or deduction instructions, and both groups received categorical induction problems that varied in inductive strength and deductive correctness. After each response, subjects were required to rate how confident they were in their responses on a 7-point scale.

The *z*ROC results of experiment 1, which also replicated Rips' main findings, are reproduced in Figure 10A. Both *H* (*P('valid' response|valid item)*) and *F* (*P('valid' response|invalid item)*) were higher in the induction than in the deduction group, indicating a more liberal response bias. The bias effect is reflected in the ROCs: points on the deduction function are clustered downward and leftward relative to the position of points on the induction function. There was also a slope difference; slope for the deduction function was higher (.84) than for the induction function (.60). Finally, analysis of *d'* revealed a sensitivity difference: *d'* was higher in the deduction than the induction group; the authors note the same conclusion was reached with the unequal-variance measure d_a. The sensitivity effect is reflected in the ROCs as well: the deduction function is higher in the space (further from the origin) than is the induction function. Of the three effects demonstrated by Heit and Rotello, the unitary view described by Rips (2001) can only predict the bias effect, i.e., that the deduction group would have a higher criterion on an argument strength dimension. It cannot account for differences in sensitivity and bias; the results therefore appear to weigh in favor of the dual-process approach.

Experiment 2 extended the initial findings by replacing the inductive strength variable with a typicality manipulation. Generally speaking, the typicality effect (Sloman, 1993; 1998) is the finding of reduced acceptance of conclusions involving atypical, relative to typical, exemplars. For instance, the argument 'All birds have property C, therefore All robins have property C' is endorsed more frequently than the argument 'All birds have property C, therefore all penguins have property C.' As can be seen in Figure 10B, the results were consistent with those of the previous experiment.

There was a main effect of typicality which did not interact with group or deductive correctness. Analysis of *H* and *F*, and visual inspection of the ROCs, again revealed more liberal responding in the induction group; sensitivity, as measured by *d'* and in terms of relative distance of the ROCs from the origin, was higher in the deduction than in the induction group; *z*ROC slope was higher in the deduction group than the induction group (.82 vs. .71). Having replicated and extended the results of their first experiment, Heit and Rotello concluded their findings could not be accounted for by a criterion shift as effects on sensitivity and slope were also obtained.

The results of these initial experiments from Heit and Rotello (2005) are important for several reasons. First, they demonstrate the power of ROC analysis as a window onto underlying processes in human reasoning; second, they illustrate the generalizability of models based on the well-established signal detection framework; third, they cross subfields to demonstrate systematization of findings which, according to some philosophers of science (e.g. Sidman, 1960), is essential in that it allows the possibility of accounting for many seemingly unrelated effects with a relatively small number of experiments. One question that remains, however, is whether the same approach used in Heit and Rotello (2005) will yield analogous findings in the area of deductive reasoning. Specifically, it is unclear whether the results obtained with categorical induction stimuli will generalize to tasks that use deductive stimuli. What can ROC curves tell us about the processes underlying performance in syllogistic reasoning tasks? Can manipulations similar to those used by Rips (2001) and Heit and Rotello (2005) be used to tease apart the contributions of system 1 and system 2 to responding in the belief bias task?

The goal of the following experiments is to determine whether the behavior of subjects in the belief bias syllogism evaluation task is best described in terms of a single- or a dual-process theory of human reasoning. The goal of experiment 1 is to replicate and extend the experiment reported by Evans and Curtis-Holmes (2005), in which a response deadline manipulation was used in an attempt to dissociate system-based responding. The goal of experiment 2 will be to determine whether the effects of induction and deduction instructions demonstrated by Rips (2001), and by Heit and Rotello (2005), generalize to syllogistic reasoning. All manipulations will proceed from the notion that two reasoning systems exist, and that conditions can be created that are more conducive to a given mode of responding.

Figure 1	Figure 2	Figure 3	Figure 4
Y-X	X-Y	Y-X	X-Y
Z-Y	Z-Y	Y-Z	Y-Z

Figure 1. The Four Syllogistic Figures.

Table 1

Design and Acceptance Rates From Evans, Barston, and Pollard (1983), Experiment 1; Adapted From Klauer et al. (2000).

Syllogism	Conclusion: Believable	Conclusion: Unbelievable
Valid	No cigarettes are inexpensive. Some addictive things are inexpensive. Therefore, some addictive things are not cigarettes. Acceptance rate: 92%	No addictive things are inexpensive. Some cigarettes are inexpensive. Therefore, some cigarettes are not addictive. Acceptance rate: 46%
Invalid	No addictive things are inexpensive. Some cigarettes are inexpensive. Therefore, some addictive things are not cigarettes. Acceptance rate: 92%	No cigarettes are inexpensive. Some addictive things are inexpensive. Therefore, some cigarettes are not addictive. Acceptance rate: 8%

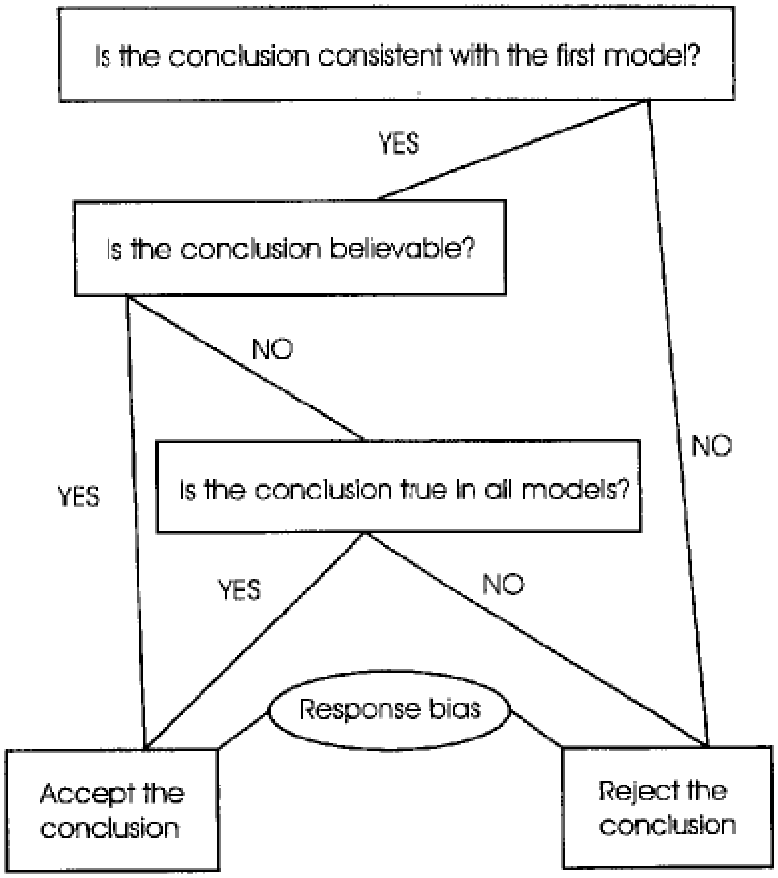


Figure 2. The Mental Models Account of Belief Bias (Adapted From Klauer et al., 2000).

Table 2

Dual-Process Theories and Their Attributes in Stanovich and West (2000)

	System 1	System 2
Dual-Process Theories:		
Sloman (1996)	associative system	rule-based system
Evans (1984;1989)	heuristic processing	analytic processing
Evans & Over (1996)	tacit thought processes	explicit thought processes
Reber (1993)	implicit cognition	explicit learning
Levinson (1995)	interactional intelligence	analytic intelligence
Epstein (1994)	experiential system	rational system
Pollock (1991)	quick and inflexible modules	intellection
Hammond (1996)	intuitive cognition	analytical cognition
Klein (1998)	recognition-primed decisions	rational choice strategy
Johnson-Laird (1983)	implicit inferences	explicit inferences
Shiffrin & Schneider (1977)	automatic processing	controlled processing
Posner & Snyder (1975)	automatic activation	conscious processing system
Properties:	associative holistic automatic relatively undemanding of cognitive capacity relatively fast acquisition by biology, exposure, and personal experience	rule-based analytic controlled demanding of cognitive capacity relatively slow acquisition by cultural and formal tuition
Task Construal	highly contextualized personalized conversational and socialized	decontextualized depersonalized asocial
Type of Intelligence Indexed:	interactional (conversational implicature)	analytic (psychometric IQ)

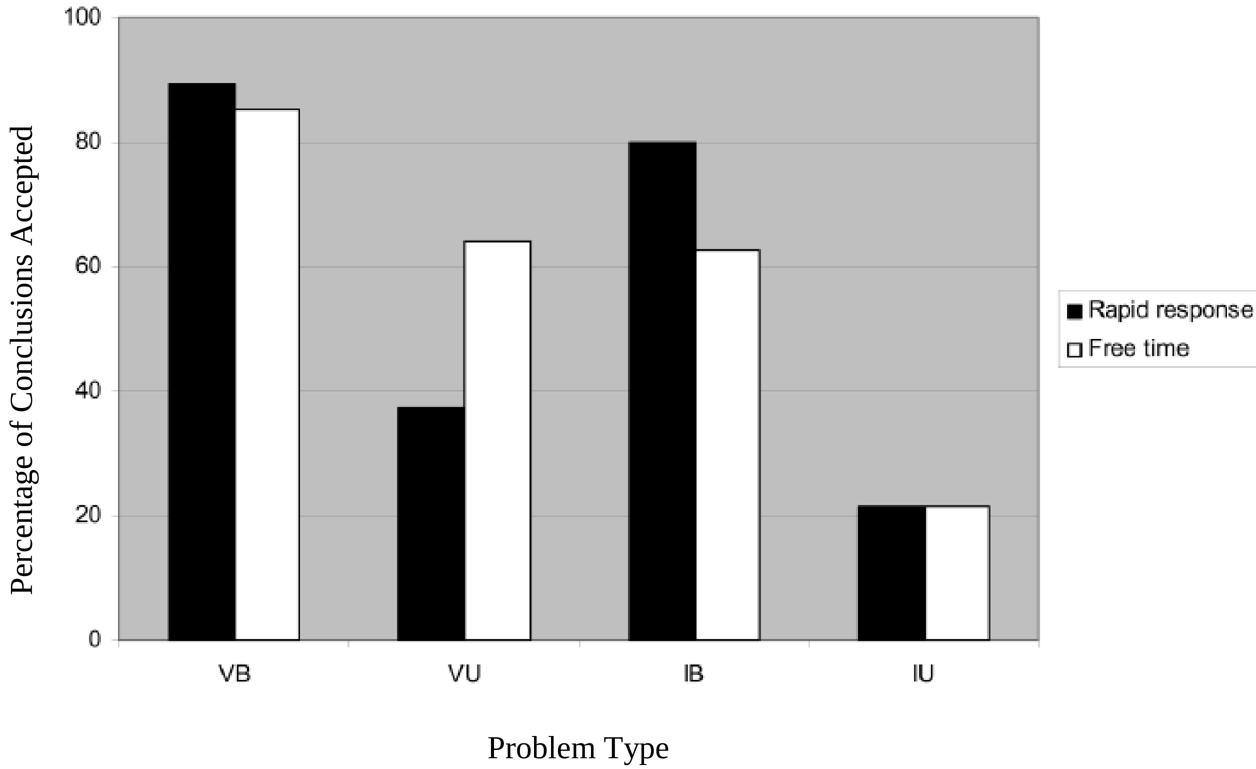


Figure 3. Percentage Acceptance as a Function of Problem Type in Evans and Curtis-Holmes (2005). V indicates valid problems, I invalid problems, B believable problems, and U unbelievable problems.

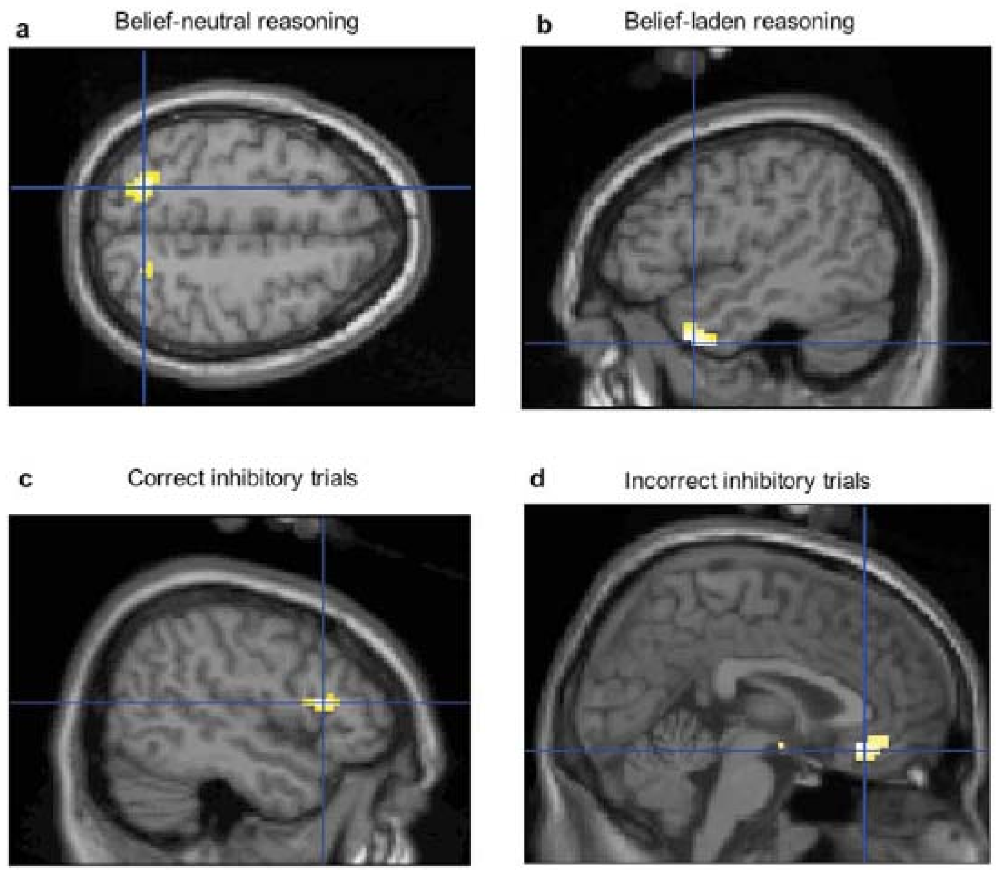


Figure 4. Neuroimaging Results From Goel and Dolan (2003). A) Belief-neutral reasoning (all responses to neutral content); scan indicates activation of the superior parietal lobule. B) Belief-laden reasoning (all responses to belief-laden content); scan indicates activation of the left pole of the middle temporal gyrus. C) Correct inhibitory trials (correct responses to valid unbelievable and invalid believable problems; scan indicates activation of right inferior prefrontal cortex. D) Incorrect inhibitory trials (incorrect responses to valid unbelievable and invalid believable problems; scan indicates activation of ventromedial prefrontal cortex.

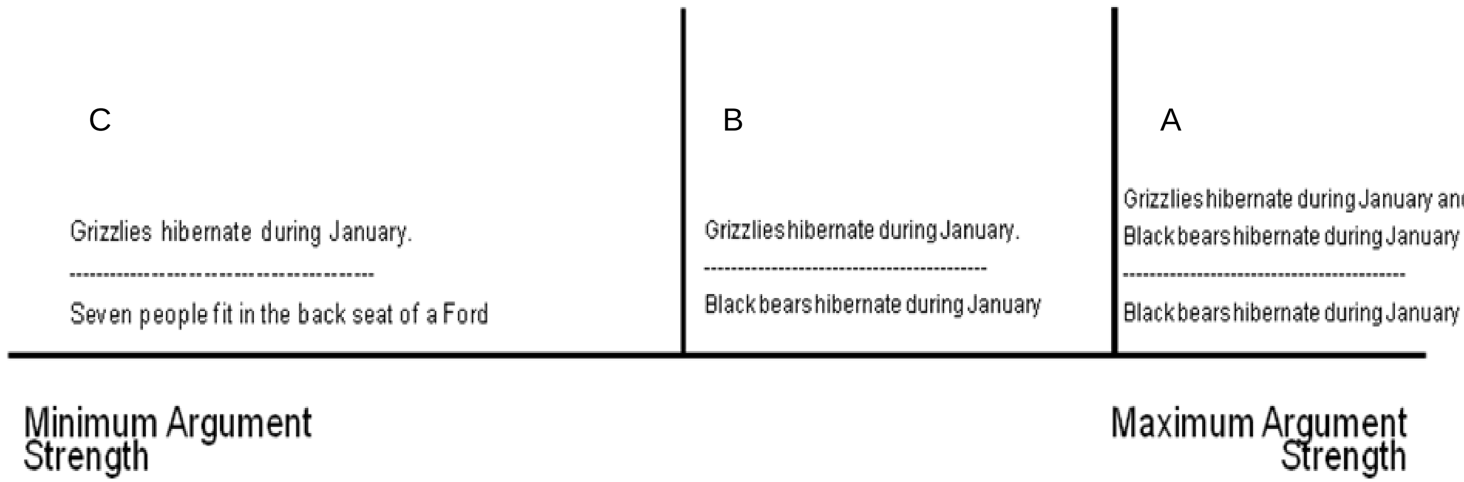


Figure 5. A One-Dimensional Account of Categorical Induction (cf. Rips, 2001). A) Believable Valid argument; B) believable invalid argument; C) unbelievable invalid argument. Vertical lines represent decision criteria, the rightmost being the more stringent position. The one-dimensional model posits acceptance and rejection of conclusions is the result of a criterion shift, and that the ordering of problems that combine believability and validity is A>B>C.

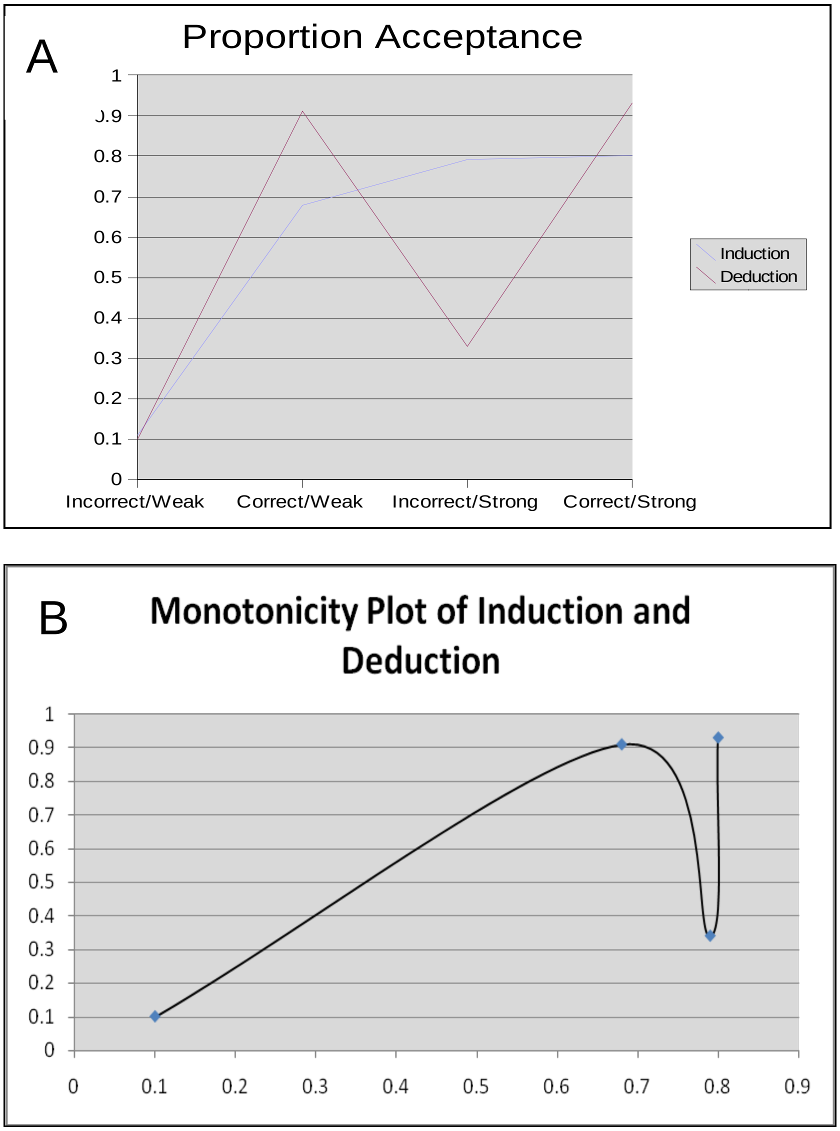


Figure 6. Results From Rips (2001). A) Proportion acceptance for induction and deduction as a function of problem type. B) Proportion acceptance for deduction (Y axis) plotted against induction (X axis); the relationship between induction and deduction changes sign as a function of stimulus, indicating a nonmonotonic relationship between the groups.

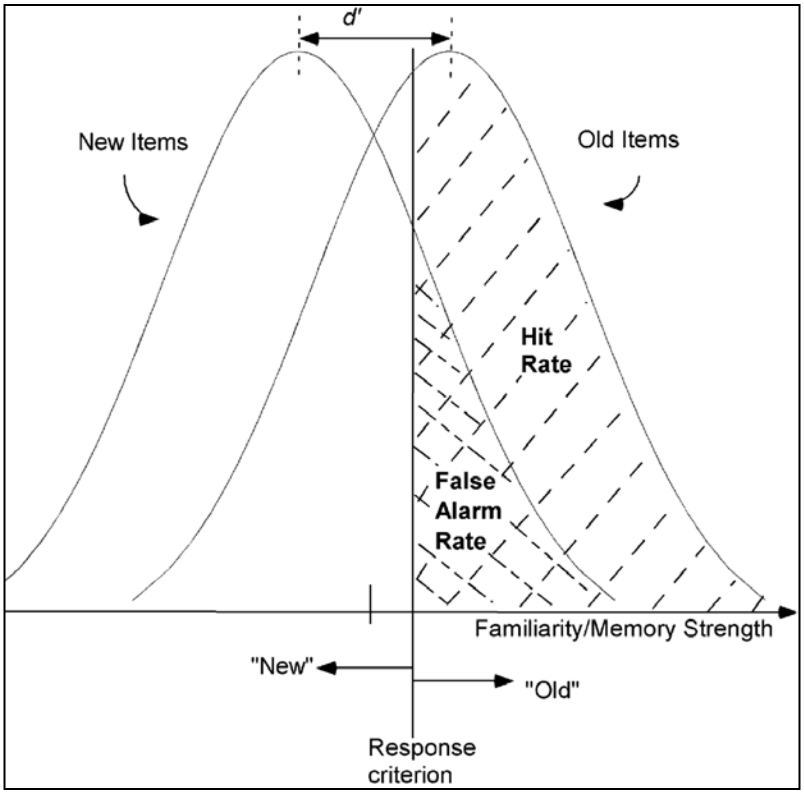
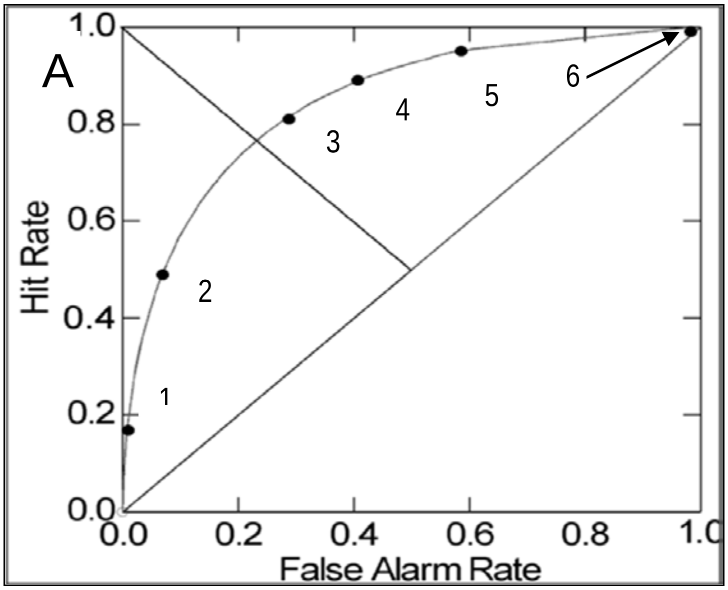


Figure 7. The Equal-Variance Signal Detection Model. The strength of items in memory is asumed to be distributed normally. The distribution of recently studied items is displaced to the right of new (lure) items, reflecting higher memory strength. Subjects differ in terms of willingness to say ‘Old’; this is modeled as a criterion dividing items into the response categories ‘Old’ and ‘New’ on the basis of their strength. The hit and false alarm rates correspond to the area under the respective old and new item distributions that falls to the right of the criterion. The distance between old and new distributions is a measure of sensitivity (*d’*) that is independent of response bias (criterion placement).



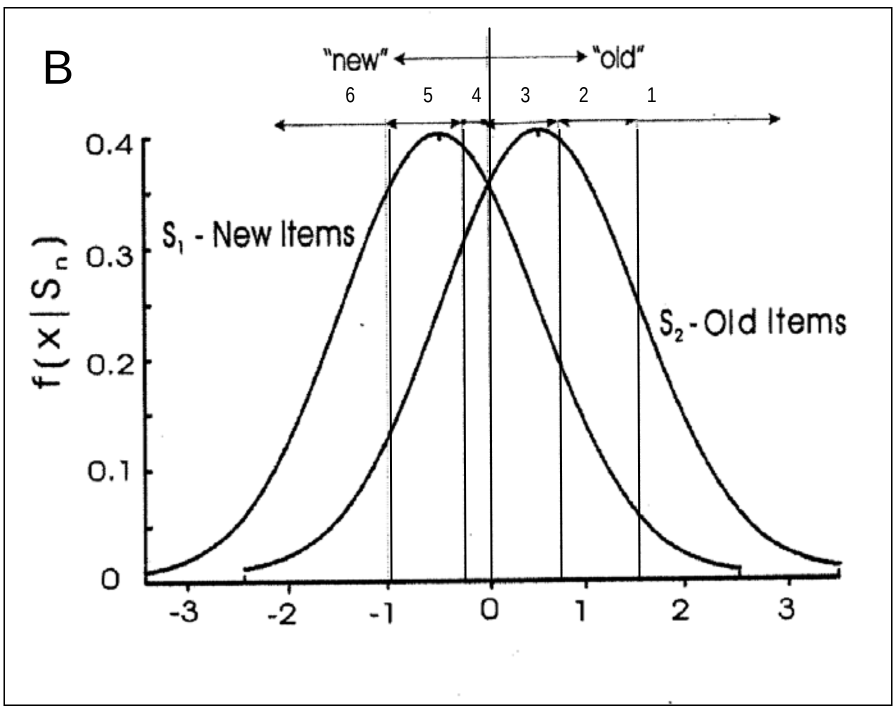


Figure 8. ROC (Receiver Operating Characteristic) Curves (Adapted From Macmillan and Creelman, 2005). A) ROCs plot hit rate (*H*) against false alarm rate (*F*) as a function of confidence. ROCs are cumulative, such that the (*F*, *H*) pair at a given point is the sum of *F* and *H* at every level of confidence up to and including that point. The distance between the ROC and the major diagonal is an index of sensitivity. The relative position of operating points on the ROC is an index of response bias; on the same curve, a '1' is a more stringent response than a '2.' B) The relationship between ratings and response bias can be understood in terms of detection theory: ratings reflect different response criteria, with a rating of '1' corresponding to the most stringent criterion in panel B.

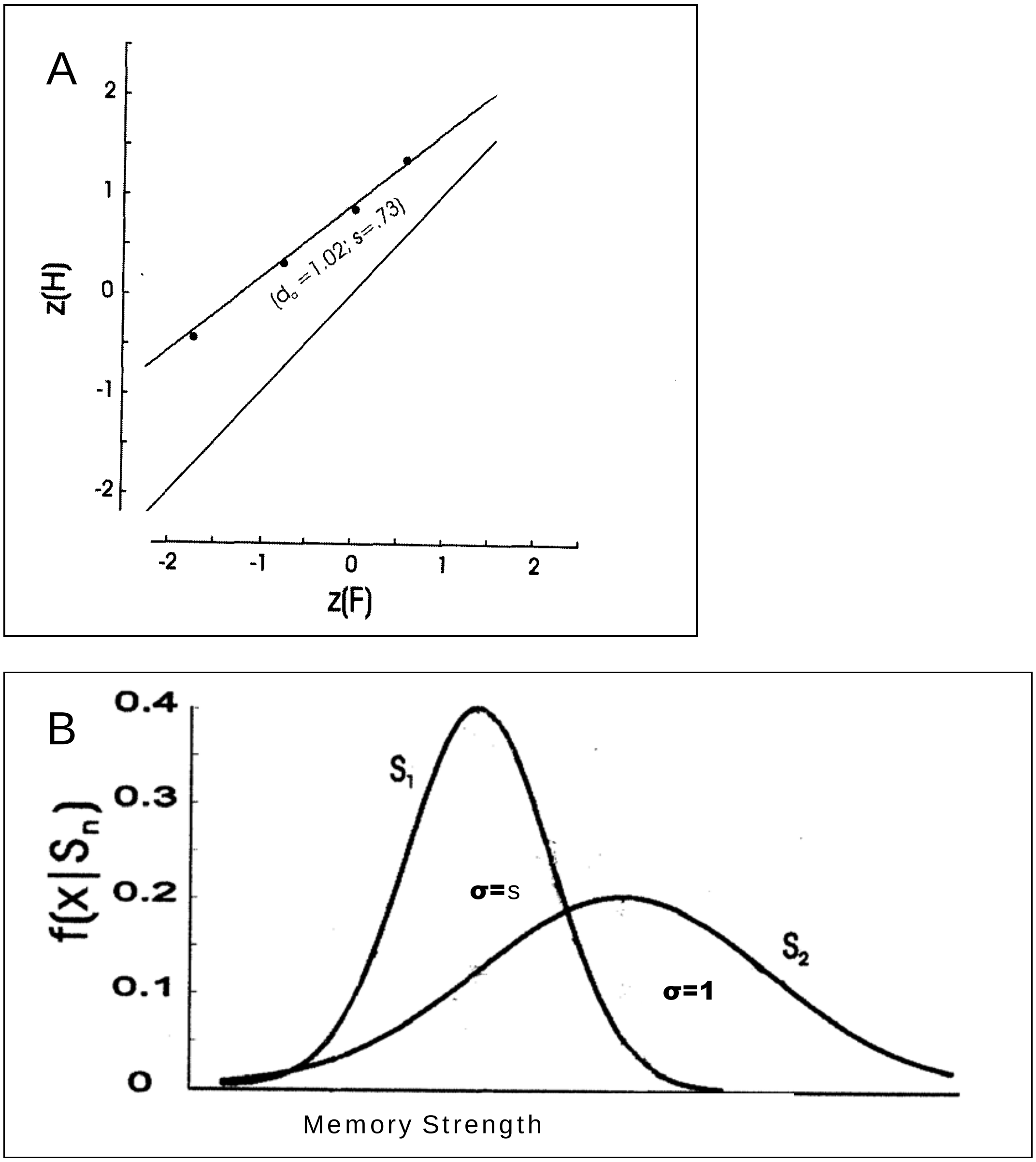


Figure 9. Unequal-Variance Detection Theory (Adapted From Macmillan and Creelman, 2005). A) Linear *z*ROC with nonunit slope; B) Unequal-variance detection theory consistent with nonunit slope in A.

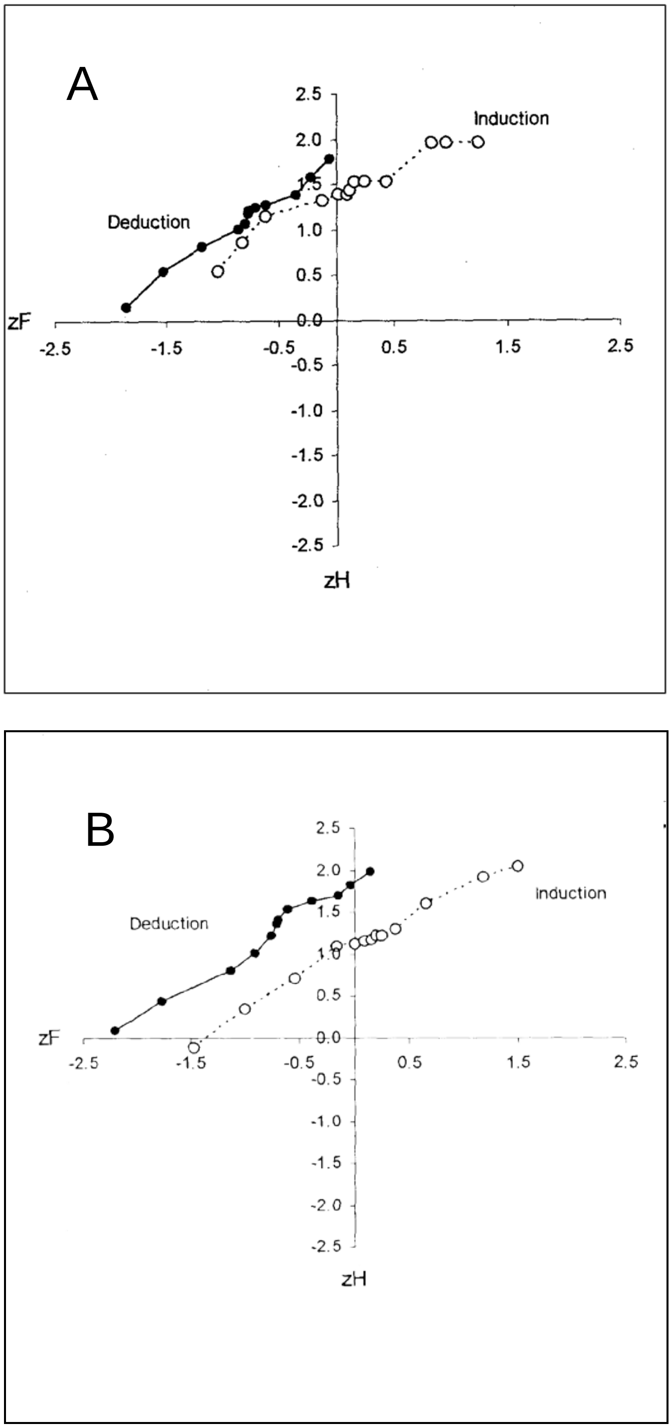


Figure 10. zROCs From Heit and Rotello (2005). A) Results from experiment 1 indicate effects of instructions on sensitivity, bias, and *z*ROC slope. B) Similar results from experiment 2.

CHAPTER II

METHOD AND RESULTS

Experiment 1

The present experiment, an extension of the study reported by Evans and Curtis-Holmes (2005), used ROC analysis to further investigate differences in system-based responding, as well as to provide information complementary to data obtained by contrasting hits and false alarms. In addition to the 10 second and unspeeded conditions of the previous study, there was a third condition in which subjects had 1 minute to respond. The inclusion of the long deadline group allowed us to assess the effect of the time limit itself. Specifically, it is possible that simply imposing a deadline is sufficient to substantially alter behavior on the whole, rather than blocking or limiting a constituent element of that behavior (i.e., system 2 reasoning). If, for example, subjects run out of time and are forced to guess or miss a deadline on one or two trials, it may inspire guessing and rapid responding on the following trials regardless of the amount of time it would actually take to reason through the problem, artifactually producing effects similar to those observed in the above study.

Method

Subjects

Experiment 1 included 119 subjects. All subjects were psychology undergraduates from the University of Massachusetts, and received course credit for their participation.

Design

Experiment 1 used a 2 x 2 x 3 mixed design. All subjects evaluated the validity of 32 syllogisms differing in logical status and believability of the conclusion; they received 8 valid believable, 8 valid unbelievable, 8 invalid believable, and 8 invalid unbelievable syllogisms. Subjects were divided into three groups: a short deadline group (n=39), in which subjects had 10 seconds to make the evaluation decision, a long deadline group (n=38), in which subjects had 1 minute to make the response, and an unspeeded group on which no time limit was imposed (n=42).

ROCs were derived by requiring each response to be followed by a confidence rating on a scale of 1 to 3, where 1 was ‘Not at all confident’ and 3 was ‘Very confident.' As the same scale was used twice (once for each response), the ROCs were plotted using 6 levels of confidence, resulting in functions with 5 points. Note that although it could be argued that confidence judgments formulated following speeded decisions may reflect the contribution of post-decisional processing, Baranski and Petrusic (1998) have demonstrated that the time taken to determine confidence under deadline conditions is unlikely to reflect the extraction of new information from the stimulus in memory.

Stimuli

Subjects evaluated 32 syllogisms. The full set of problems was comprised of two subsets, each containing equal numbers of valid and invalid problems. Set A included 8 structures that fully control for atmosphere, conversion, and figural effects. As in Evans, Barston, and Pollard (1983), atmosphere and conversion were controlled in Set A by using both invalid and valid forms of problems using the logically convertible premise quantifiers ‘Some’ and ‘No’, and conclusion quantifier ‘Some…are not’ which is favored by the premise atmosphere of ‘Some’ and ‘No.’ Figures 2-4 were used, and figural effects were controlled for by presenting conclusions for figure 4 in directions both preferred and nonpreferred by the bias, at both levels of validity. Set B contained 8 additional structures for which atmosphere and figure were controlled as in Set A, but each problem allowed illicit conversion. Although the premise quantifiers were convertible, the effect of conversion for these particular problems is unlikely to produce artifactual belief bias effects as, unlike the original problem set examined by conversion theorists (e.g. Revlin, Leirer, Yopp, & Yopp, 1980), the converted versions of each problem lead to the same response. Premise quantifiers ‘All’, ‘No’, and ‘Some…are not’ were used, in figures 2 and 3; like Set A, all problems used ‘Some…are not’ conclusions (for the actual structures, see Appendix C).

The 8 problems in each set were repeated twice each, once with a believable conclusion and once with an unbelievable conclusion, yielding the full set of 32 problems. Problem content for 13 problems was taken from a previous study by Morley et al. (2004); new content was used for the remaining 19 problems. All sets of content were randomly assigned to the 32 problem structures. For the new content, conclusion believability was rated previously by a group of 59 psychology undergraduates at the University of Massachusetts in Amherst, using a scale from 1 to 5 where a 1 corresponded to ‘unbelievable’, a 3 corresponded to ‘neutral’, and a 5 corresponded to ‘believable.’ The most extreme ratings were then selected to construct the present set of stimuli. The conclusions, along with means and standard deviations, are presented in Appendix B. All content was chosen such that conclusions related a statement about a category-exemplar relationship between subject and predicate terms. In order to minimize the effects of premise believability, subject and predicate terms were linked via an esoteric middle term (e.g. ‘No sculptors are hammerkops/Some hammerkops are not artists’).

Content was counterbalanced such that it appeared in both believable and unbelievable, and both valid and invalid structures. Between subjects, modulation of belief status was accomplished by reversing the order of assignment of words to the subject and predicate positions. In other words, for each subject that received the conclusion ‘Some spiders are not insects’, an equal number received the conclusion ‘Some insects are not spiders’, while no subject received both. Further, for each of the 16 structures the actual believable or unbelievable content was also varied. Counterbalancing thus yielded 4 subsets of 32 problems.

Finally, practice problems used in experiment 1 (see *Procedure*) included esoteric predicate terms in order to create belief-neutral conclusions (e.g. ‘Some cowboys are theurgists’).

Procedure

All subjects were tested individually and were seated approximately two feet in front of a computer monitor. During an initial preparation phase, deduction instructions were read to the subject who was then shown three neutral example problems (two valid problems and one invalid) and asked to reiterate in his or her own words the meaning of the terms valid and invalid. Instructions and preparation materials are listed in Appendix D.

The procedure for the unspeeded group was as follows: upon completion of the preparation phase subjects received a welcome message; once the message had been read the subject advanced the experiment via key-press. Next, deduction instructions were displayed, followed by the message “Before we start the experiment, let’s try a few practice trials. Press any key to begin practice.” Once subjects advanced the message, a syllogism was presented, followed by the response options 'Not valid’ or 'Valid.' Subjects indicated their response via key-press (F for 'Not valid’ or J for 'Valid’). Once the evaluation response was made, a new screen containing the question “How confident are you in this judgment?” appeared, along with a description (“1 = Not at all confident, 2 = Moderately confident, 3 = Very confident”) and the instructions “Press key: 1 2 3.” Once the confidence response was made, the process repeated for the remaining 4 syllogisms. Practice problems were concrete but contained neutral content (see *Stimuli*). Upon termination of the practice session, there was an intermission message informing subjects that they could take a quick break and to advance to the experimental trials via key-press. Experimental trials proceeded in the same manner as the practice trials, but contained a new set of 32 belief-laden syllogisms (see *Stimuli*). Order of presentation for the 5 practice problems and 32 experimental problems was completely randomized for each subject.

In the short deadline group, the same procedure was followed as in the unspeeded group, but with the following changes. All on-screen instructions were augmented to explain the deadline procedure. In addition, the practice trials of the unspeeded group were replaced by a series of 5 trials using the deadline procedure (see Appendix E for deadline instructions). The procedure followed closely that of Evans and Curtis-Holmes (2005). On a given deadline trial, only the premises (and the line) of the syllogism were presented for the first 5 seconds of the trial, followed by presentation of the conclusion below the premises for an additional 5 seconds. A time clock appeared at the start of each trial, counting backward from 10 seconds in 1 second intervals. If and when the clock reached the final second, the timer was replaced by the message "make a decision now." Subjects who failed to make a decision before the termination of the final second were advanced to the next trial and no response was recorded for the missed trial. Once the evaluation decision was made, subjects were advanced to a new screen asking for a confidence rating. Confidence ratings were unspeeded, and this was indicated in the instructions.

Following completion of the training phase, subjects received the intermission message indicating completion, as in the unspeeded group. This was followed by the first experimental trial, which involved the same procedure as in the training trials and repeated for the full set of 32 syllogisms.

The procedure for the long deadline group was the same as that of the short deadline group, with two exceptions. One exception is that the premises and conclusion were presented simultaneously, in order to render conditions comparable to the unspeeded group, which follows the more traditional design of belief bias experiments. These conditions were appropriate for assessing the effect of imposing a long deadline relative to standard conditions in which no deadline is imposed. The other exception is that the time clock counted backward from 60 seconds, which was also reflected in the instructions.

Results

Proportion of Deadlines Missed

One limitation of the study reported by Evans and Curtis-Holmes (2005) is that the effect of missed trials in the 10 second condition is not known. In order to assess the effect in the present study, for each subject the proportion of trials in which the deadline was missed (P(M)) was calculated. With one outlier excluded (P(M) = .31)), the data were normally distributed with mean = .08 (approximately 3 trials missed) and SD = .06 (approximately 2 trials missed). P(M) was not influenced by whether problems were believable or not ($t(37) = .896$, $p = .376$) or by whether they were valid or not ($t(37) = 0.000$, $p > .05$). The data were then split at the median, with subjects below the median assigned to a 'low missed' group and those above the median assigned to a 'high missed' group. A 2 (group: high vs. low) x 2 (logical status) x 2 (believability) mixed ANOVA indicated no interaction between the effects of logic or belief with group on P(M) ($F(1, 36) = 0.000$, $MSE = .007$, $p > .05$, and $F(1, 36) = .050$, $MSE = .013$, $p > .05$, respectively). Finally, the analyses reported below were conducted both with and without the high missed group, and both with and without the outlier for whom P(M) = .31. As none of the conclusions reached by analysis of the full sample were affected by either of these variables, it was concluded that subjects were randomly missing relatively small numbers of trials. The analyses reported below were conducted on the full sample.

Hits and False Alarms

The proportion of conclusions accepted was analyzed using a 2 x 2 x 3 mixed ANOVA with logic and belief as within-subjects factors and group as a between-subjects factor. Interactions were examined using paired comparisons, which were Bonferroni-corrected in order to minimize the contribution of familywise error.

Results for hits and false alarms (summarized in Table 3) imply the standard belief bias effect. First, there was a main effect of logic, indicating greater acceptance rates for valid than invalid problems, $F(1,116) = 184.968$, $MSE = .044$, $p<.001$. Second, there was a main effect of belief, indicating greater acceptance rates for believable than unbelievable problems, $F(1,116) = 73.126$, $MSE = .043$, $p<.001$. Third, there was a logic x belief interaction, indicating a greater effect of belief for invalid than for valid problems, $F(1,116) = 12.402$, $MSE = .026$, $p<.01$.

Group and logical status also interacted, $F(2,116) = 8.615$, $MSE = .044$, $p<.001$. Paired comparisons revealed the logic index (H – F) was larger in the unspeeded than the 10 second group, $t(116) = 3.494$, $p<.01$, and the same relation held for the 60 second relative to the 10 second group, $t(116) = 3.687$, $p<.01$. The 60 second and unspeeded groups did not differ in sensitivity to logical status, $t(116) = .308$, $p>.05$. There was also an interaction between group and belief, $F(2, 116) = 7.164$, $MSE = .043$, $p<.01$; the belief index (P("Valid"|Believable) – P("Valid"|Unbelievable)) was marginally larger in the 10 second than the unspeeded group, $t(116) = 2.423$, $p = .051$, and the same relation held for the 10 second relative to the 60 second group, $t(116) = 3.715$, $p<.01$. The 60 second and unspeeded groups did not differ in sensitivity to belief, $t(116) = 1.397$, $p>.05$.

Finally, no 3-way interaction was obtained, indicating that the interaction between logic and belief was comparable across the three groups, $F(2, 116) = 1.264$, $MSE = .026$, $p = .286$. Though this aspect of the data was not consistent with the results of Evans and Curtis-Holmes (2001), it should be noted that conflicting results have been previously reported by Shynkaruk and Thompson (2006), implying that the problem might not lie with dual-process theory, but with theorists' interpretations of the interaction in general (see *General Discussion*). With the exception of the null belief x logic x group result, the results for H and F are thus consistent with the data reported by Evans and Curtis-Holmes (2001), and suggest that imposing a deadline, in and of itself, does not substantially alter responding.

ROC Analyses

Testing for apparent differences in ROCs using detection theory parameters requires correcting for H=1 and F=0, and as recognition experiments typically avoid these levels of responding, corrections can be made that do not substantially impact the results of significance tests. However, the belief bias effect in the present experiment produced unusually large numbers of potential corrections (22% of believable hits and false alarms; 13% of unbelievable hits and false alarms). As a result of this issue, parameter tests were not included in this analysis and it was necessary instead to directly compare the functions. ROCs are plotted in Figures 11-13. Gray lines indicate the upper and lower bounds of the 95% confidence interval for each (bold) group ROC. Confidence intervals were obtained by bootstrapping 2000 samples from the individual data and selecting group ROCs falling at the 2.5th and 97.5th percentiles of the resulting distribution. To more closely examine the effects of logic and belief, two types of ROCs were plotted. Logic ROCs plot hits against false alarms, where the hit rate is defined as P("Valid"|Valid) and the false alarm rate is defined as P("Valid"|Invalid). Belief ROCs plot hits against false alarms, where the hit rate is defined as P("Valid"|Believable) and the false alarm rate is defined as P("Valid"|Unbelievable).

The main effect of logic is reflected in plot 11A; the chance line is below the lower bound of the 95% CI for the logic ROC. Plot 11B reflects the main effect of belief; the chance line is below the lower bound of the CI for the belief ROC. Interestingly, the robust belief x logic interaction obtained with the measure H-F was not replicated in the ROCs (Figure 11C). The confidence interval of the logic ROC for believable problems overlaps with the confidence interval of the unbelievable logic ROC. Though decisive conclusions regarding the appropriate sensitivity statistic cannot be drawn without fitting models assumed by those statistics to the observed ROCs, the absence of an effect in ROC height indicates H-F is not an appropriate measure of sensitivity for this task.

To further clarify the measurement discrepancy, ROCs implied by H - F have been superimposed on the observed ROCs in Figure 11D. The ‘interaction index’ used in studies of belief bias is a contrast of the logic index (H - F) for unbelievable and believable problems (HU - FU - HB + FB); this is the equivalent of a contrast of f(x) for the lines intersecting the (F, H) pair at the midpoint of the corresponding ROCs, for a single value of x. The midpoints of the ROCs yield the group average of F and H for problems that are believable (.63, .84) and unbelievable (.42, .73), which can also be obtained using the averages from Table 3. The difference in f(x) between these functions for any single point along the x-axis is always .10, the value of the interaction index, though the difference in f(x) for the observed ROCs at a single value of x is not constant at all values of x. The ROCs implied by H - F do not appear to map on to the observed ROCs and are thus unlikely to provide an accurate measurement of the interaction between logic and belief.

Figures 11C-D also indicate a marked shift in response bias: the position of operating points is upward and rightward for the believable relative to the unbelievable ROC. This is another reflection of the belief bias apparent in 11B, which implies furthermore that the conditions under which belief bias is typically obtained are also just the conditions under which H-F is likely to lead one to erroneous conclusions (i.e. when response biases differ; Macmillan & Creelman, 2005; Rotello, Masson, & Verde, 2008). The slope of the zROC also appears to be greater for believable logic (.97) than unbelievable logic (.81); this is consistent with the notion that the argument strength distribution containing conflict problems (where belief and logic disagree) will be more variable than the distribution containing facilitatory problems (where belief and logic agree). For the believable ROC, one might expect the invalid item distribution to be more variable than the valid item distribution, raising the slope relative to the unbelievable ROC, which might be expected to reflect less variability in the invalid than in the valid distribution, yielding slope = .81.

The interaction between the effects of logic and group is shown in Figures 12A-C. While the confidence intervals for the logic ROCs of the 60 second and unspeeded groups overlap (12C), the function for each group is higher in x-y space than the function for the 10 second group (12A-B), indicating relatively lower sensitivity to logical status in the 10 second group.

The interaction between belief and group is illustrated in Figures 13A-C. While the confidence intervals for the belief ROCs of the 60 second and unspeeded groups overlap (13C), the function for the 10 second group is higher in the space than the function for each of the other groups (13A-B). In all cases, zROC slope exceeds 1, indicating the variance differential for belief ROCs is opposite that of the logic ROC (slope = .89 for logic, 1.19 for belief). Again, this can be readily explained in terms of the inclusion of conflict items in the distributions the ROCs are thought to reflect. Specifically, if the variance of valid unbelievable argument strength is great enough to exceed that of the other three argument types, then one would expect relatively high slope estimates when those arguments are treated as comprising part of the noise distribution (belief ROCs), while slope would be correspondingly lower when the same arguments are treated as comprising part of the signal distribution (logic ROCs).

Perhaps the most interesting result in terms of the form of the ROCs can be seen in the belief ROC for the 10 second group. In the logic ROCs of the 60 second and unspeeded groups, where system 2 reasoning should predominate, there is a suggestion of two-piece linearity that seems to be entirely absent in the more curvilinear belief ROC of the 10 second group. If one assumes the contributions of system 1 and system 2 are reflected in the logic and belief indices, then the appropriate ROC-based equivalents for assessing these factors would be the logic and belief functions. Furthermore, if system 1 responding does predominate in the 10 second group, and the belief ROC to some degree reflects that contribution, then the ROCs suggest system 1 might be better characterized as a continuous process like that assumed by the detection theory model, while the contribution of system 2 might be better approximated by a threshold model assuming two or more discrete states. Until further work is directed at fitting such models to these data, however, the only conclusion that can be reached is that the belief and logic ROCs, when they are substantially above the chance line, appear to differ in form.

Discussion

The results for ROCs and H - F are generally in accordance, and the findings of increases in logic- and decreases in belief-based responding with extra time reported by Evans and Curtis-Holmes (2005) were replicated. The inverse relationship between belief- and logic-based responding over time cannot be explained by a single-process account, such as the one detailed by Rips (2001). The one-process view could only have predicted a reduction in overall accuracy (i.e. only one latent variable, with its own time course, would have been manipulated) and/or a more liberal or conservative criterion with increased time pressure. The results are consistent with the notion that imposing a short deadline blocks the contribution of system 2, thereby eliminating opportunities for analytic processes to override the conclusions of erroneous, fast-acting heuristic processes. Imposing a (long) deadline or missing deadlines does not appear to substantially affect how subjects respond.

The finding of an increase in the interaction between belief and logic with time was not replicated, however. A related finding was that the robust belief x logic interaction obtained by contrasting H and F was not replicated in the ROCs. The form of the ROCs suggests a model different from the one implied by H - F (a straight line of unit slope), which raises questions regarding the statistical and theoretical significance of the interaction originally reported by Evans, Barston, and Pollard (1983). However, the use of a relatively stringent instruction phase, as well as the existence of conflicting results regarding the interaction and its interpretation (see *General Discussion*), indicate more work is needed before any decisive conclusions about the interaction and its theoretical implications can be reached. Finally, the finding of apparent differences in the form of the belief and logic ROCs supports dual-process theory, though firm conclusions must await the application of models which assume fundamentally different processes.

Experiment 2

The goal of experiment 2 was to determine whether the effects of induction and deduction instructions demonstrated by Rips (2001), and by Heit and Rotello (2005), would generalize to syllogistic reasoning. In addition to the belief bias task used in experiment 1, a control condition was included in which both instruction groups solved a block of abstract syllogisms. Though effects of induction and deduction were not expected for the abstract block, the question remained as to whether subjects in the belief bias task would rely on two processes to respond to a given stimulus, or whether they would rely on a single process to combine information from two stimulus *attributes*, i.e., the logical status and believability of a given problem. We entertained the possibility of effects of induction and deduction for abstract problems, assuming such results would provide convincing evidence in favor of dual-process theory, while acknowledging that the lack of an effect would still be consistent with the commonsense notion that abstract problems can only be solved using system 2 reasoning.

The inclusion of the abstract block served two further purposes. First and foremost, it was hoped that the form of the abstract ROC might be useful to future research by providing information about the nature of logic-based processing, in addition to serving as a comparator for belief-laden ROCs. Secondly, the inclusion of abstract and belief-laden ROCs in experiment 2 allowed an additional question to be addressed: could the solution of abstract syllogisms have an effect on system-based responding? There is some evidence in the developmental literature that suggests it could (Hawkins, Pea, Glick, & Scribner, 1984; Markovits & Vachon, 1989; 1990). For instance, Hawkins et al. (1984) presented 4-5 year old children with syllogisms containing either fantasy or realistic content (content invoking knowledge about the world) and found that when fantasy problems were solved prior to realistic problems, performance overall was better than when the order was reversed. In addition, coded 'justification' data revealed that the fantasy-first group provided more theoretical (logically deductive) justifications on both problem types than did the other groups. Hawkins et al. concluded that “A theoretical or abstract attitude toward the verbal problems appears to have been made possible because the fantasy problems were constituted of premises isolated from practical knowledge,” (p. 592). It was uncertain whether this effect would generalize to adults or whether an 'abstract attitude' might reflect system 2 processing, but as one of the main questions of experiment 2 required the collection of abstract ROCs, it was relatively easy to test for this effect in adults.

Method

Subjects

Experiment 2 included 122 subjects. All subjects were psychology undergraduates from the University of Massachusetts, and received course credit for their participation.

Design

Experiment 2 used a 2 x 2 x 2 x 2 mixed design. Subjects evaluated the same set of 32 syllogisms as in experiment 1. Subjects were divided into two groups; 61 subjects received induction instructions and the other 61 received deduction instructions (see Appendix A for actual instructions). Subjects in each group were subdivided into an abstract and a belief-only group. The abstract group (n = 59) evaluated a block of 16 syllogisms containing letters in place of words for the subject, predicate, and middle terms; following the abstract block, subjects received a block of 32 stimuli containing words as in experiment 1. The belief only group (n = 63) received only the block of 32 concrete syllogisms.

Stimuli

Details regarding syllogisms used in experiment 2 are the same as those for experiment 1, with the exception that all problem structures were presented three times to half of the subjects: once in believable, unbelievable, and abstract forms. Abstract versions of the 16 structures were created by randomly selecting a set of 24 letters from the alphabet, each of which was randomly assigned two times to the terms of the syllogisms, with the constraints that no two letters shared a problem more than once, and that no letter was repeated in a given problem.

Procedure

All subjects were tested individually and were seated approximately two feet in front of a computer monitor. Subjects were randomly assigned to one of four conditions: deduction/abstract (n = 31), deduction/belief only (n = 30), induction/abstract (n = 28), induction/belief only (n = 33). There was an initial preparation phase during which deduction or induction instructions were read to the subject who was then shown three neutral example problems (two valid/strong problems and one invalid/not strong) and asked to reiterate in his/her own words the meaning of the terms Valid/Invalid or Strong/Not strong. All subjects were then asked to complete three practice problems and indicate the confidence of their responses on a scale of 1 to 3 where a 1 corresponded to low confidence in the response and a 3 corresponded to high confidence. Practice materials are listed in Appendix F.

Subjects in the abstract groups evaluated a block of 16 abstract syllogisms the terms of which contained letters of the alphabet in place of words. The procedure in this case was similar to the unspeeded condition of experiment 1: instructions (in this case, induction or deduction) were presented on-screen, followed by a syllogism and the response option “Not valid (F) or Valid (J),” followed by a confidence rating on a scale of 1 to 3, where a 1 indicated “Not at all confident” and a 3 indicated “Very confident.” There was then an optional, untimed rest interval; this was followed by a second block of 32 experimental trials using a similar set of concrete syllogisms (see *Stimuli* for details). The sequence of events for the second block of trials was the same as that of the previous block.

Subjects in the belief only groups underwent the same procedure as the abstract group with the exception that they only completed one block of 32 stimuli; these were the same syllogisms as those used in the second block of the abstract condition.

Results

Hits and False Alarms

Hits and false alarms are summarized in Tables 3-4. For abstract problems, the logic index (H – F) did not differ as a function of instructions, $t(57) = 1.629$, $p > .05$, consistent with the notion that if subjects in the two groups were responding on the basis of two reasoning systems, they would have been constrained to rely on system 2 to solve these problems.

The proportion of conclusions accepted was analyzed initially using a 2 x 2 x 2 x 2 mixed ANOVA with conclusion believability and logical status as within subjects factors and instructions (deduction vs. induction) and sequence (abstract vs. belief-only) as between subjects factors. As no effects or interactions were obtained in relation to the sequence variable, the data were collapsed across this factor and the remaining variables were analyzed using a 2 x 2 x 2 mixed ANOVA.

Two results were obtained. First, there was a main effect of logic, indicating greater acceptance rates for valid than for invalid problems, $F(1,120) = 262.599$, $MSE = .051$, $p<.001$. Second, there was a main effect of belief, indicating greater acceptance rates for believable than for unbelievable problems, $F(1,120) = 44.977$, $MSE = .037$, $p<.001$. The interaction between logic and belief was not significant, though there was a trend in the expected direction, $F(1,120) = 2.779$, $MSE = .021$, $p=.098$. No other main effects or interactions approached significance.

ROC Analyses

As in experiment 1, the number of potential corrections for H=1 and F=0 was quite high (25% of believable hits and false alarms, 14% of unbelievable hits and false alarms), so it was again necessary to directly compare the ROCs. Accordingly, the resampling procedure of experiment 1 was applied to the present dataset. ROCs are plotted in Figures 14-16; the gray lines refer to the upper and lower bounds of the 95% confidence interval for each (bold) group ROC.

Figure 14A indicates a main effect of logic; the chance line is below the lower bound of the 95% CI for the logic ROC. 14B indicates a main effect of belief; the chance line is below the lower bound of the 95% CI for the belief ROC. Although, as in experiment 1, the slope of the belief ROC appears to exceed that of the logic ROC (1.09 vs. .83, respectively), in this case for every comparison belief ROCs were very low in x-y space. As all belief ROCs are likely to be constrained in form by their proximity to the major diagonal, any conclusions based on apparent differences in slope or shape must be weighed with caution.

The data in figure 14C are consistent with the results for H-F; the interaction between logic and belief was not obtained in the present experiment. Additionally, the slope difference indicated in experiment 1 for believable vs. unbelievable logic ROCs was not apparent in these functions. This could reflect the fact that in the present experiment (as well as in the unspeeded group of experiment 1) belief-based responding was attenuated relative to logic-based responding (14A-14B). From the standpoint of detection theory, this relatively greater reliance on logic could have reduced the effect of conflict problems on the argument strength distributions, such that any effect on zROC slope would have been the result primarily of logical validity. Also, as in experiment 1, operating points on the believable ROC are shifted upward and rightward relative to the corresponding points on the unbelievable function. This indicates relatively greater willingness to accept believable conclusions, and is a further reflection of the effect of belief illustrated in 14B.

Figure 15A supports the null effect of instructions on the processing of abstract problems inferred from H-F. The functions are comparable in height, slope (.70 for deduction, .81 for induction), and position of operating points. Though the shape of the functions appears to differ somewhat, it must be noted in this respect that the sample size was relatively small (n=28 for induction, n=31 for deduction), as was the number of arguments (8 valid, 8 invalid). These conditions are likely to introduce more noise into the ROCs than in previous comparisons.

Figures 15B-C agree with the results for H-F in indicating no effect of instructions on the effects of logic or belief; the induction and deduction CIs for logic ROCs overlap (15B), as do the CIs for belief ROCs (15C).

An interesting finding in experiment 2 was that the ROCs for abstract and belief-laden stimuli (16A) appear to differ very little in form, height, position of operating points, and zROC slope (.83 for the logic ROC, .75 for the abstract ROC). This indicates, in line with the conclusion suggested by the comparison of induction and deduction on abstract problems, that regardless of whether one expects processing to differ on the basis of induction vs. deduction (Rips, 2001) or conclusion believability (Evans and Curtis-Holmes, 2005), it does not appear that separate reasoning systems contribute to the processing of information on a single dimension in this task.

Finally, Figure 16B indicates, as in experiment 1, that processing of logic and belief may differ fundamentally. The abstract and belief ROCs appear to differ in slope (.75 and 1.09, respectively), as well as shape, with the abstract ROC exhibiting the same two-piece linearity present in prior comparisons of logic ROCs. This seems to be entirely absent in the belief ROC, as in experiment 1. As mentioned above, however, the belief ROCs of the present experiment tend to approximate the major diagonal, rendering any such comparison potentially misleading.

Discussion

The present experiment failed to demonstrate an effect of instructions stressing induction or deduction on performance in a syllogistic reasoning task. While on the surface this may suggest the findings of Rips (2001) and Heit and Rotello (2005) are to some extent task-specific, there are a number of potential reasons for the lack of an effect that remain to be explored. For instance, syllogistic reasoning is a fairly difficult task, and effects of belief in the present study, though statistically significant, are small in comparison to the effects of logic. This may indicate that factors inherent in the syllogistic task bias subjects toward deductive reasoning at the outset. If this is so, the question remains as to why belief bias is so prevalent in syllogistic reasoning. It may be that, though related, the processing distinction delineated by Rips (2001) does not map onto the one drawn by Evans and Curtis-Holmes (2005). Another possibility is that the emphasis on instructions, which required subjects to repeat back the stated reasons for conclusions to be considered valid or strong, may have had an effect, though essentially the same in both groups. That is, actually engaging subjects in the arguments during the instruction phase may have biased all subjects toward deductive behavior, despite a superficial difference in whether such proctored deductions were labeled 'valid' or 'strong.' This would explain both the lack of an instructional effect in the present study, as well as the unexpected interaction results for both experiments reported in this writing.

Additionally, it should be noted that a null effect of instructions for syllogisms despite such effects for categorical induction may actually be consistent with the notion that separate inductive and deductive reasoning systems exist, assuming the traditional distinction between deductive and inductive arguments is a product of something more than academic or pedagogical tradition.

Finally, the prior solution of abstract syllogisms did not have an effect on subsequent reasoning with belief-laden material. This is not overly surprising, as the effect reported by Hawkins et al. (1989) was limited to coded justification data and has never been documented with adults. Additionally, Chen & Daehler (2000) demonstrated that, in order to instantiate transfer in insight problem solving by way of a prior analogy, it was be necessary to use analogies that were very similar to the target problem, or that were actually generated by the subject; a single, dissimilar analogy did not affect performance relative to controls given irrelevant material in place of an analogy. This suggests that for transfer in syllogistic reasoning to occur, subjects might profit most from training that makes explicit the similarity between the logical structure of abstract and belief-laden examples, as well as from training in the translation of belief-laden problems into their abstract equivalents.

Table 3

Proportion of Conclusions Accepted by Group and Problem Type, Experiment 1 and 2.

	Experiment 1			Experiment 2	
Problem Type	10 seconds	60 seconds	Unspeeded	Deduction	Induction
Valid	.72	.85	.80	.81	.85
Invalid	.57	.53	.49	.47	.53
Believable	.77	.73	.72	.70	.74
Unbelievable	.51	.65	.57	.58	.63
Valid Believable	.81	.86	.86	.87	.88
Valid Unbelievable	.62	.84	.73	.75	.81
Invalid Believable	.73	.60	.57	.53	.60
Invalid Unbelievable	.40	.45	.40	.40	.45
Logic Index	.30	.65	.62	.69	.64
Belief Index	.52	.16	.29	.25	.22
Interaction Index	.14	.13	.04	.01	.08

Logic index = P("Valid"|Valid) – P("Valid"|Invalid); belief index = P("Valid"|Believable) – P("Valid"|Unbelievable); interaction index = logic index(Unbelievable) – logic index(Believable).

Table 4

Proportion of Abstract Conclusions Accepted in Experiment 2, by Group.

Problem Type	Deduction	Induction
Valid	.80	.83
Invalid	.53	.44
Valid - Invalid	.27	.39

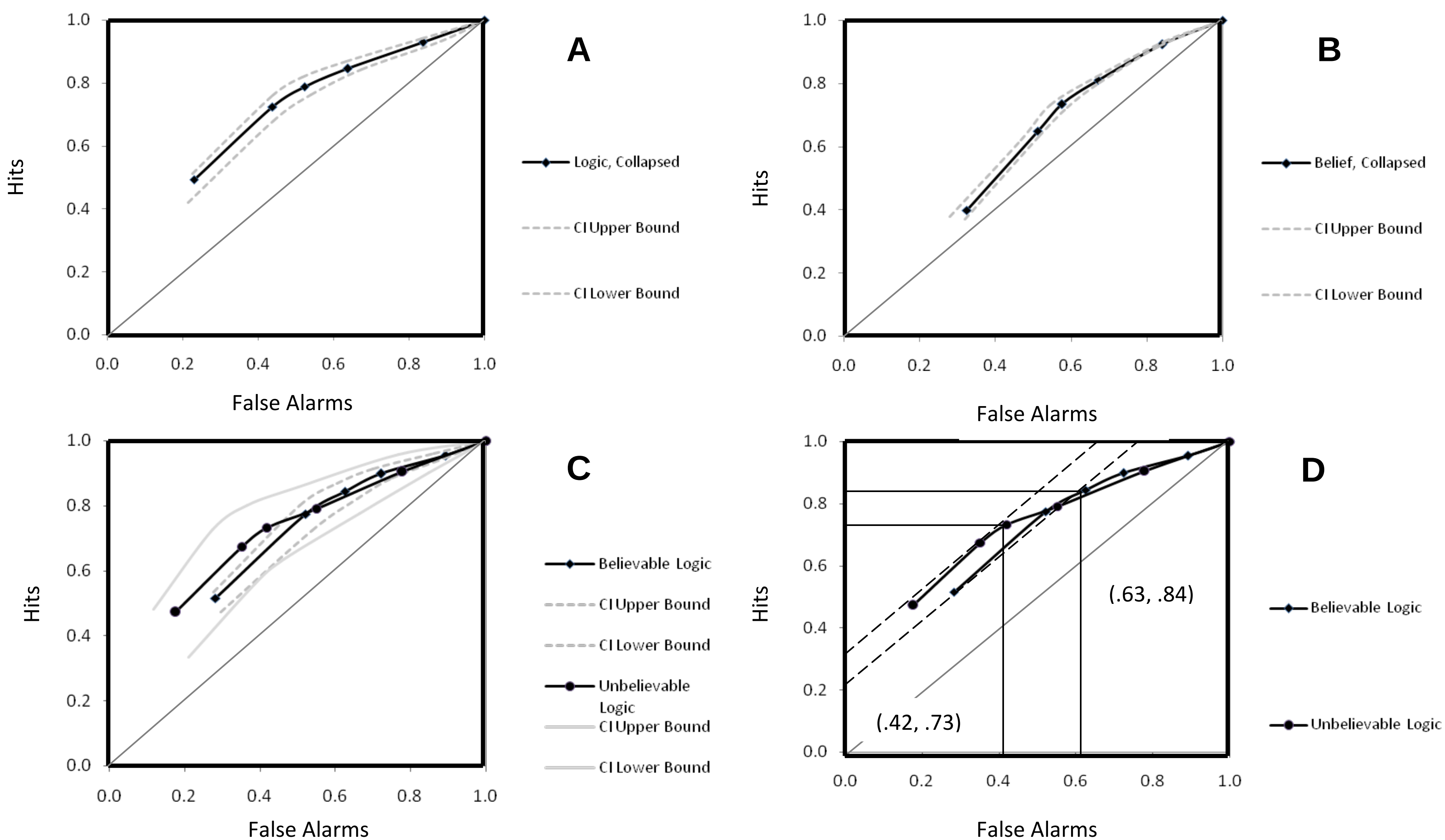


Figure 11. ROCs From Experiment 1. Gray lines indicate the upper and lower bounds of the 95 % confidence intervals for each bold ROC. A) Logic ROC, collapsed across groups. Hits = P("Valid"|Valid), false alarms = P("Valid"|Invalid). B) Belief ROC, collapsed across groups. Hits = P("Valid"|Believable), false alarms = P("Valid"|Unbelievable). C) Logic ROCs for syllogisms with believable and unbelievable conclusions. D) 11C with ROCs implied by H - F superimposed (dashed lines). Interaction Index = .73 - .42 - .84 + .63 = .10.

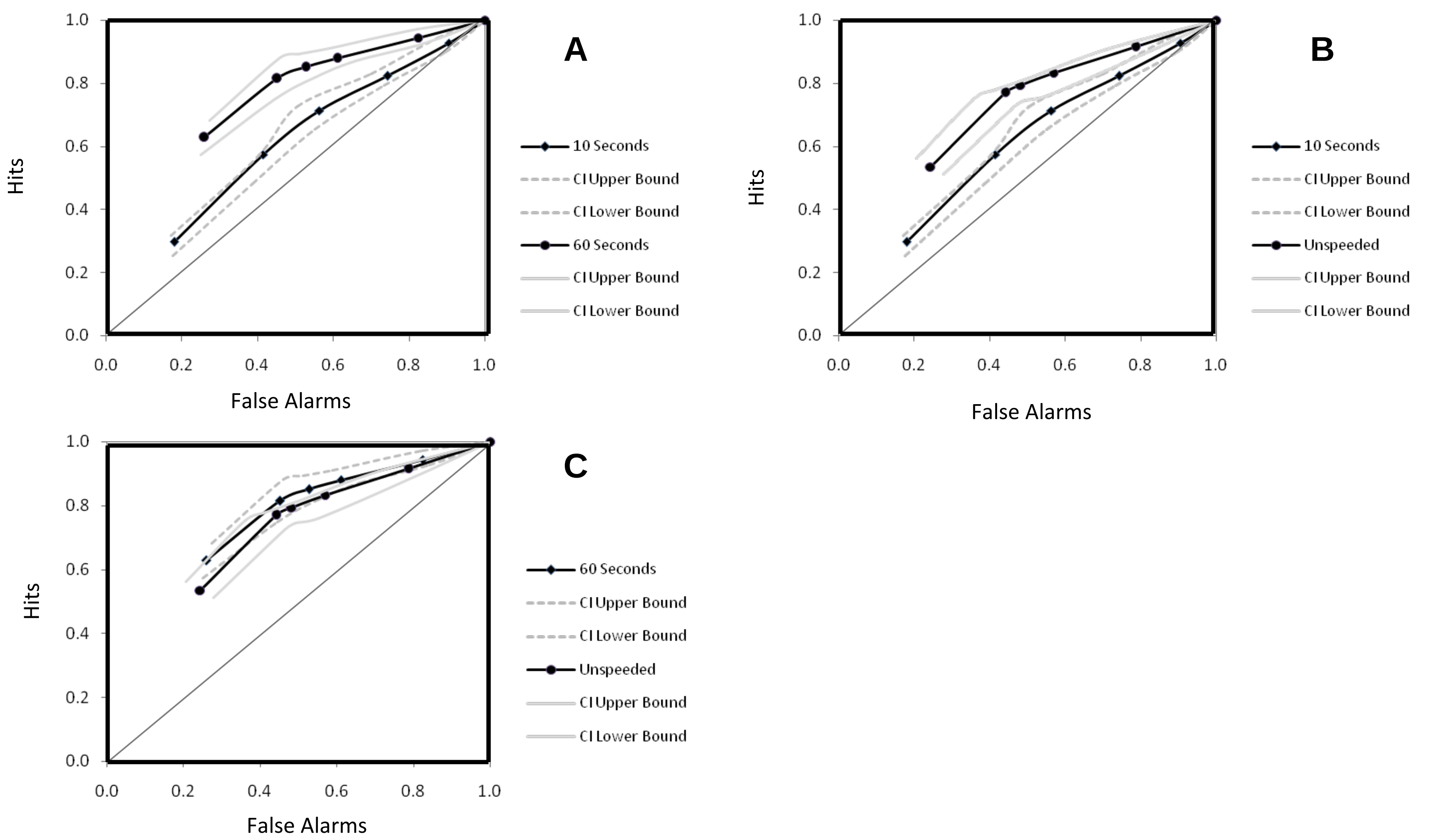


Figure 12. Logic ROCs From Experiment 1, by Group.

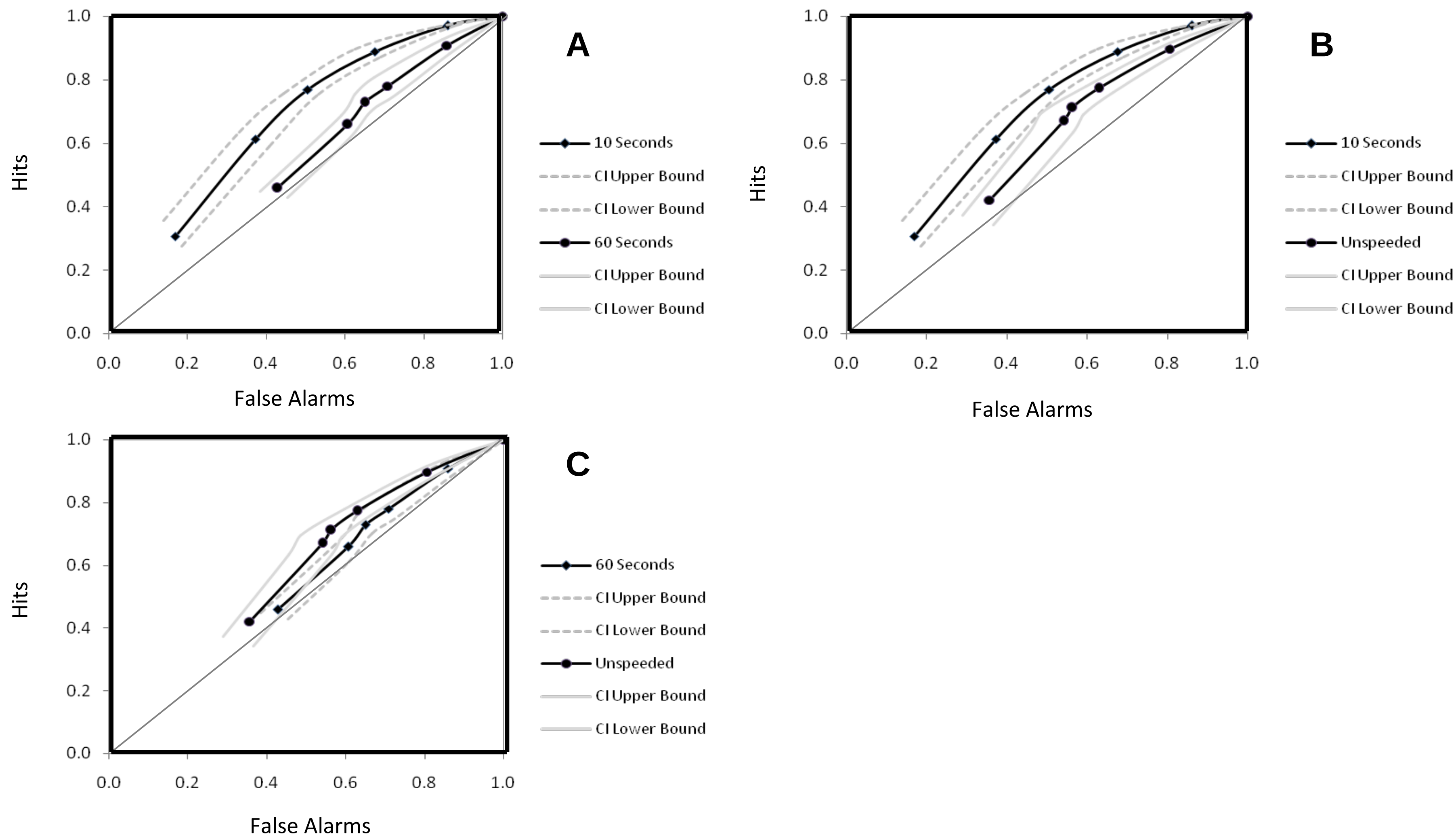


Figure 13. Belief ROCs From Experiment 1, by Group.

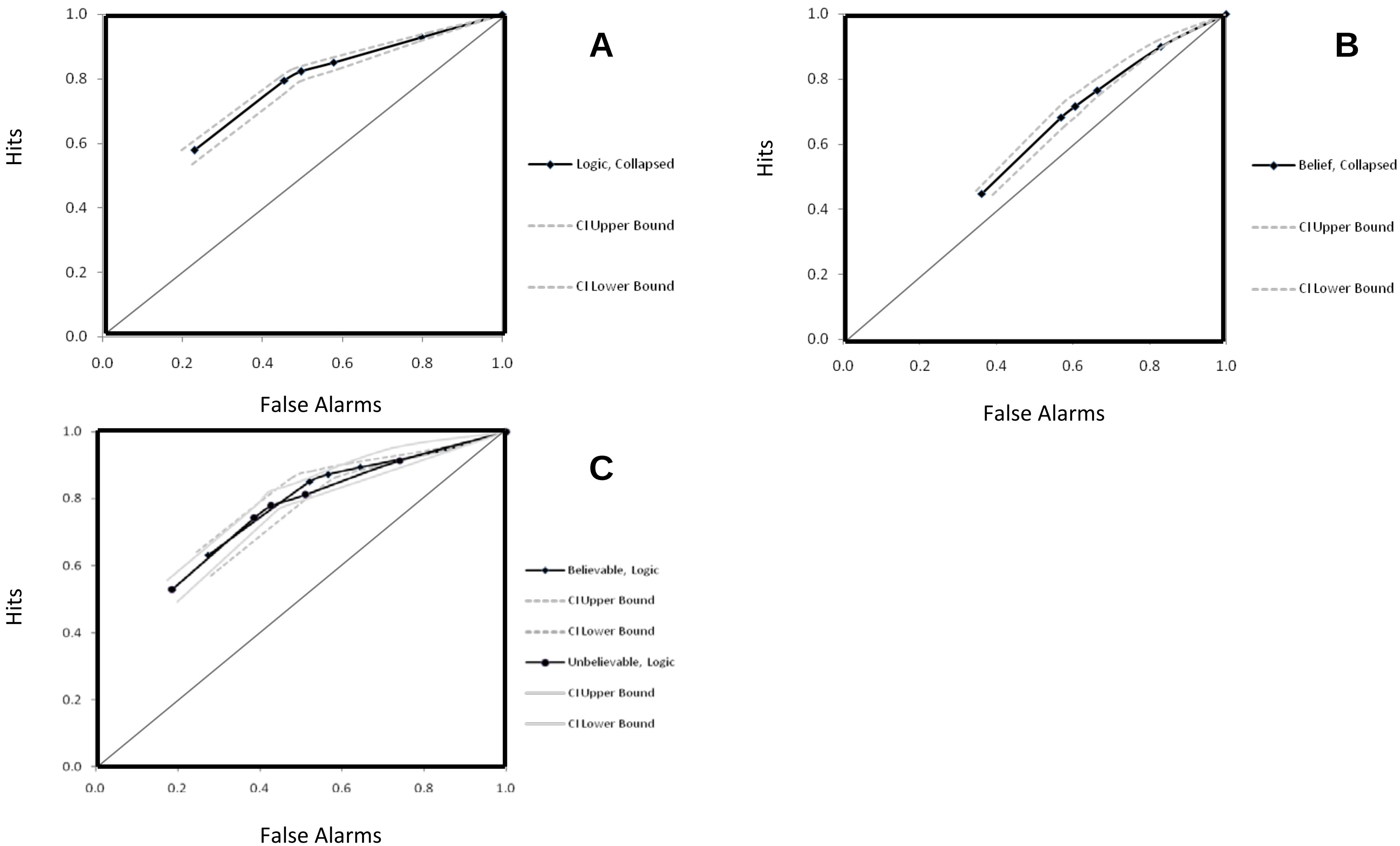


Figure 14. ROCs From Experiment 2. Gray lines indicate the upper and lower bounds of the 95 % confidence intervals for each bold ROC. A) Logic ROC, collapsed across groups. Hits = P("Valid"|Valid), false alarms = P("Valid"|Invalid). B) Belief ROC, collapsed across groups. Hits = P("Valid"|Believable), false alarms = P("Valid"|Unbelievable). C) Logic ROCs for syllogisms with believable and unbelievable conclusions.

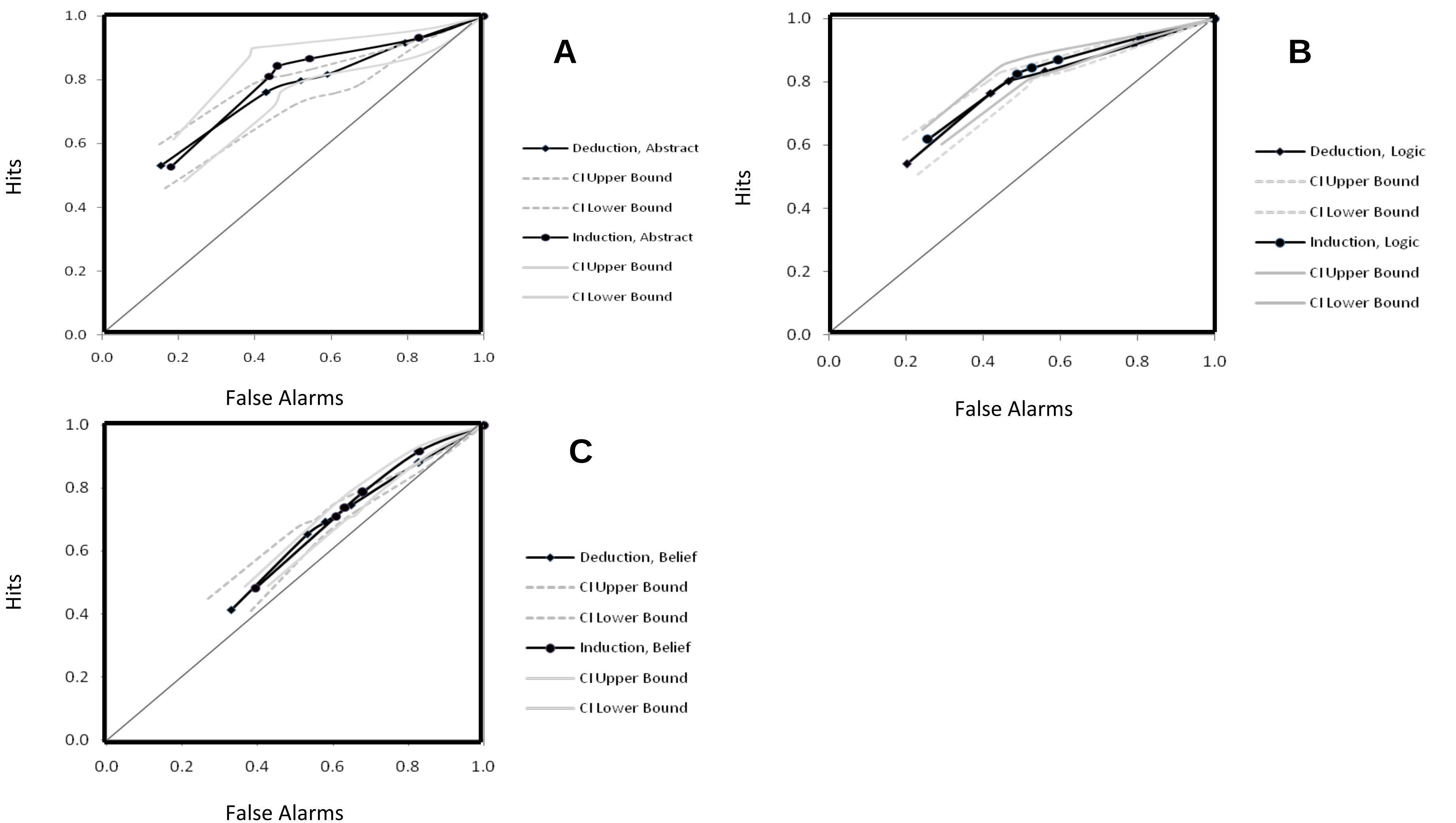


Figure 15. Abstract and Belief-Laden ROCs, by Group. A) ROCs for abstract syllogisms, by group. B) Logic ROCs, by group. C) Belief ROCs, by group.

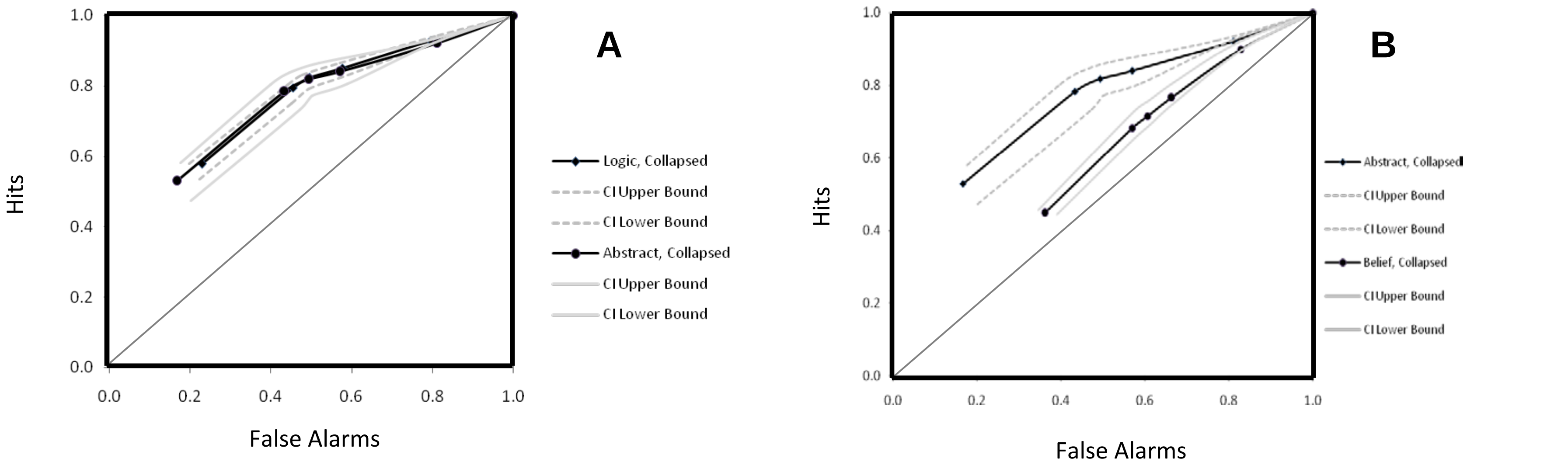


Figure 16. Abstract and Belief-Laden ROCs, Collapsed. A) A comparison of logic and abstract ROCs, collapsed over groups. B) A comparison of abstract and belief ROCs, collapsed over groups.

CHAPTER III

GENERAL DISCUSSION

While experiment 1 provided support for the 'heuristic-analytic' theory of Evans and Curtis-Holmes (2005), which proposes subjects apply system 1 and system 2 processes to the evaluation of syllogisms, subjects do not appear to process syllogisms inductively when instructed to do so. Though the results of experiment 2 could be interpreted as imposing a limitation on the conclusions reached by Rips (2001) and Heit and Rotello (2005), there are several reasons for caution in doing so, all of which indicate the need for further work investigating differences inherent in inductive and deductive arguments, as well as modes of reasoning.

For example, it is possible that the very act of parsing the premises of syllogisms may require deductive processes, e.g. to represent quantification and/or to form an initial representation or model linking subject and predicate via the middle term. This is not the case for the types of arguments employed by Rips (2001) and Heit and Rotello (2005). Consider, for instance, the conjunction elimination argument type employed by both studies:

Jill does D and Jill does R

Jill does D. (10)

The evaluation of (10) does not require subjects to do anything more than 'look up' the conclusion in the premises. Thus, while Rips' subjects were (nominally) asked whether the conclusion was 'necessarily true' they may have been (functionally) asked whether the conclusion 'restates information printed above the line.' In the case of the syllogism (4) below, the conclusion does not appear in print, and can only be determined after some effort is made on the part of the reasoner to restate or represent the problem in a way that is not explicit in the stimulus (i.e. it is not stated in print).

All X are Y

No Y are Z

No Z are X. (4)

The notion that the processing of premises of propositional arguments (¼ of the arguments used by Rips were of this sort) is fundamentally different from the processing of syllogistic premises is supported by conflicting findings in the literature regarding the effects of premise believability (Thompson, 1996; Torrens, Thompson, & Cramer, 1999; Klauer, Musch, & Naumer, 2000; Markovits & Schroyens, 2007). Thompson (1996) presented subjects with valid and invalid propositional arguments that varied in premise and conclusion believability and found that subjects were more likely to accept believable conclusions when the premises were also believable than when they were not, with the size of the effect being similar for valid and invalid problems. For arguments with believable and unbelievable conclusions neutral and unbelievable premises did not differentially affect acceptance rates for either valid or invalid arguments. In contrast, a follow-up experiment by Klauer, Musch, & Naumer (2000; experiment 8) found, for a set of syllogisms containing two critical invalid arguments, that subjects were more likely to accept unbelievable conclusions when the premises were unbelievable than when they were neutral. It is unclear why Klauer et al. obtained conflicting results, but until more work is directed toward understanding how differences in reasoning tasks such as these interact with premise-based reasoning, firm conclusions about the generality of the inductive/deductive theory may remain elusive.

The present results also suggest that syllogistic arguments may be substantially more difficult than inductive or categorical arguments in general, and this added difficulty could have reduced the effectiveness of induction instructions. The relative difficulty of syllogistic reasoning is implied in the sensitivity indices: accuracy for deduction subjects in experiment 2 appears to be somewhat lower than in the comparable condition of Heit and Rotello's experiment 2 (d' = .95 vs. d' = 2.11, respectively). Additionally, it is worth pointing out that syllogistic reasoning is commonly thought to be demanding of working memory resources, with several studies demonstrating a positive relationship between accuracy and memory span in syllogistic reasoning in conclusion production (Copeland & Radvansky, 2004) and evaluation tasks (Quayle & Ball, 2000), for both younger and older adults (Gilinsky & Judd, 1994). It is not clear how different are the constraints on memory for induction and syllogistic reasoning, but it would be quite surprising if arguments like (10) were nearly as demanding in this respect as arguments like (4). Assuming syllogistic arguments do indeed place greater demands on working memory resources, no representational difference is actually necessary to account for the present results. Subjects may simply become so involved in the processing demands of the relatively complex syllogistic arguments that they fail to follow or simply forget the instructions detailed at the outset.

Another important difference between the argument types is that the ones used by Rips (2001) and Heit and Rotello (2005) were largely generics (arguments without explicit quantification). As stated in the introduction, a major source of difficulty in syllogistic reasoning stems from the effects of quantification (Begg & Denny, 1969; Dickstein, 1975). Though the present study attempted to control the effects of atmosphere and conversion, more recent research has indicated subjects may have particular trouble with the logical interpretation of quantifiers ‘Some’ and ‘Some..are not.’ That is, though logically 'Some X are not Y' is consistent with both 'Some X are Y' and 'No X are Y', the interpretation of such statements by subjects appears to reflect conversational interpretations that are more or less consistent with Grice's (1975/2000) maxim of informativeness. Specifically, it appears that 'Some X are not Y' and 'Some X are Y' are both taken to imply 'Some X are not Y (and some X are Y)' (Roberts, Newstead, & Griggs, 2001; Schmidt & Thompson, 2008). In support of this idea, Schmidt & Thompson (2008) have shown that, when logically clarified quantifiers such as 'At least some' are used in place of the more ambiguous 'Some', errors in reasoning with neutral syllogisms are substantially reduced. As the ‘Some’ and ‘Some…are not’ quantifiers were used extensively in the present study (and in many previous studies of belief bias), it may be that quantification has added a potential source of difficulty to syllogistic arguments that is not present in categorical or propositional ones. If this is true, simply adding quantification that nonetheless maintains the logical status of the categorical induction arguments used in Heit and Rotello's experiment 2 should lower performance overall for both induction and deduction subjects, and, critically, the acceptance rates for conflict items in the two groups should even out.

Finally, it is possible that the instructions would actually have been effective but that in both cases the preparation phase, in which subjects were asked to repeat back the meaning of the words 'valid' or 'strong', may have pushed all subjects toward a more deductive approach. Anecdotally at least, subjects tended when probed about the meaning of key terms to refer back to the premises of the argument rather than to the stated definitions of 'valid' or 'strong'. The inclusion of example problems in the instructions, coupled with the probe question, may have led subjects to approach the problems deductively in both conditions. This could explain the failure to obtain the usual interaction between logic and belief, especially if subjects in both groups were led to discover the principle of logical necessity, which seems to be such an important part of the deductive approach.

Evidence for the effectiveness of instruction in logical necessity is mixed. An early study by Dickstein (1981) demonstrated that errors in syllogistic reasoning could be substantially reduced by including instructional emphasis on logical necessity, but the effect was only obtained for invalid problems, with additional instructions actually worsening performance on valid problems. An additional concern is that the results may not be informative for belief bias experiments in that the materials employed were abstract and the task was not conclusion evaluation, but 5AFC. More recently, Newstead et al. (1992; experiment 5) contrasted the effects of standard instructions and instructions augmented to explain logical necessity on reasoning in the syllogism evaluation task with belief-laden material. In the standard group, effects of belief, logic, and an interaction were obtained; in the augmented group, the belief index was reduced, the logic index was increased, and the interaction did not reach significance. Though not explicit in the discussion of their results, the authors may have demonstrated a reduction in the belief x logic interaction by emphasizing analytic processing (more on this later). A follow-up study by Evans et al. (1994) failed to replicate the effect reported by Newstead et al.;

nonetheless, their experiment 3 did demonstrate a statistically null belief bias effect, and a significant reduction in the belief x logic interaction, when an extended (and very complex) set of augmented instructions was used. The complex instructions of Evans et al. (1994) did not contain additional passages relating explicitly to logical necessity, however, leading the authors to conclude factors other than necessity may have contributed to both their results and those of Newstead et al., though the authors did not attempt to pinpoint a specific aspect of the complex instructions that could have produced the result. Thus, though the theory of misinterpreted necessity may not provide an exhaustive account of belief bias (see *Introduction*), it appears that an understanding of the necessity concept may, under some circumstances, influence the extent to which subjects engage in deductive behavior when reasoning with syllogisms.

If the emphasis on instructions in the present experiment actually drew attention to syllogistic premises and away from the nominal definition of induction and deduction, it could have led subjects to understand the issue of logical necessity in both conditions when encountering the third, indeterminately invalid example (see *Preparation Instructions*, *Appendix D*). This would explain the lack of an interaction in the present experiment, as well the differences between experiment 1 and the study reported by Evans and Curtis-Holmes (2005) (see below). Additionally, if instruction in logical necessity were shown to have stable effects on reasoning with relatively complex syllogistic arguments, it could also raise questions regarding the interpretation of the results reported by Rips (2001). In essence, the difference in conclusion acceptance rates for induction relative to deduction may not have been due to emphasis on 'induction' per se, but rather to an absence of the emphasis on logical necessity which appeared in the deduction instructions.

The results of experiment 1 are mostly consistent with the data from Evans and Curtis-Holmes (2005), and more generally, with the predictions of heuristic-analytic theory articulated by Evans (2006), in which fast-acting system 1 processes are said to onset relatively early, to supply information that is operated upon subsequently by the more logically-oriented system 2. When subjects are constrained to respond within 5 seconds of presentation of belief-laden conclusions that logically relate premise information, the beliefs cued by those conclusions appear to dominate responding. When subjects are given extra or unlimited time to respond, the influence of belief is reduced and the influence of logic is increased, in line with the idea that system 2 may intervene at a relatively later stage to override responses cued by system 1 (Evans, 2006).

For the effects of logic and belief, the ROC results were consistent with results obtained using the index H-F. Logic ROCs were higher in x-y space when subjects were allowed extra or unlimited time to respond, while belief ROCs were higher in the space when subjects were constrained to make speeded decisions.

An unexpected result was the finding of no difference in the height of believable and unbelievable logic ROCs, despite a robust belief x logic interaction in the analysis of H-F. This can easily be explained as a consequence of inadvertently (and incorrectly) assuming the threshold model implied by H-F. More specifically, H-F and related statistics such as proportion correct (.5*(H+1-F)) assume that sensitivity is being measured independently of response bias, i.e. that the ROC plotting H against F as a function of levels of willingness to say 'Valid' will be linear, with slope = 1. Though the logic ROCs obtained in the present experiments appear to exhibit linearity, it doesn't appear that the data would be best described by a regression line, let alone a line of unit slope. Whether and to what extent this sort of error has contributed to research regarding the belief x logic interaction is unclear. There are, however, important reasons to withhold conclusions regarding the apparent reliability of H-F as an index of the interaction.

First, no conclusive statements regarding the shortcomings of H-F can be made until the assumptions of threshold models are actually evaluated by fitting them to the data. Neither threshold nor any competing models were fit to the results of the present experiment, though in light of this issue and the ROC results (to be discussed below), model selection appears to be an important next step in research regarding the time course of syllogistic reasoning.

Second, the interaction results in both experiments conflict with prior findings in the belief bias literature. Though there was no belief x logic x group interaction apparent in the results of experiment 1, a post-hoc test confirmed that the interaction index did not differ from zero for the unspeeded group, $t(41) = .860$, $p = .395$. Similarly, no interaction was obtained for either group in experiment 2.

Third, the post-hoc test suggests further that the interaction in the present study (nonsignificantly) decreased with an increase in time available for reasoning, which is diametrically in opposition to the result reported by Evans and Curtis-Holmes (2005). This is also in opposition to the interpretation of the interaction by mental models and selective processing theorists (Ball et al., 2006; Oakhill, Johnson-Laird, & Garnham, 1989; Polk & Newell, 1995) which assumes it is a product of logical, rather than heuristic, processes. The present results may suggest analytic processing has made a surprisingly profound contribution in the 10 second group of the present experiment. It may also mean that the interaction is actually a product of heuristic, not analytic, processes. Interestingly, one of the few studies besides that of Evans and Curtis-Holmes to compare conclusion evaluation under a deadline of 10 seconds with performance under a longer deadline (60 seconds) also found an interaction between logic and belief in the shorter deadline condition (Shynkaruk & Thompson, 2006). Further, when subjects were given extra time (1 minute) to reconsider their responses, the interaction did not decrease despite an increase in the logic index, similar to the corresponding between-subjects result of the present study. The authors replicated this pattern in a second experiment. In their general discussion, Shynkaruk and Thompson stated that "..one must conclude that the interaction is not due to formal reasoning processes but, rather, arises from the application of fast and simple heuristics, which can be applied in about 10 sec., " (p. 630). Note that the same implication also follows from the instruction results of Newstead et al. (1992) and Evans et al. (1994), in which training in logical necessity reduced the interaction. Taken together, these findings seem to converge on the notion that the belief x logic interaction is not a product of analytic processes as has previously been assumed, but is due rather to heuristic processes, in agreement with Shynkaruk and Thompson (2006).

On the other hand, the interaction is likely dependent to some degree on effects of both logic and belief, and shortcomings of the present study in terms of the size of these effects may be responsible for reducing the interaction. Specifically, though the magnitude of belief and logic effects appears to be comparable in the present experiments, the logic effect was in both the 10 second and unspeeded groups substantially larger than in the comparable conditions of Evans and Curtis-Holmes' study (for the 10 second groups: *d* = .81 and .53, respectively; for the unspeeded groups: *d* = 1.32 and 1.01). The belief effect was also smaller in the present study than in the earlier one, and the difference appeared to be more pronounced when subjects had to make speeded decisions (for the 10 second groups: *d* = 1.30 and 2.27, respectively; for the unspeeded groups: *d* = .69 and .91).

It is unclear why the studies should differ in this way. One possibility is that the increase in logic-based responding is a result of the preparation instructions, mentioned earlier. If subjects adopted a more deductive approach to the problems as a result of the question probe, and particularly if subjects were led to discover the principle of logical necessity, it could have produced the differences in effect size between the two studies, as well as reducing the effect of induction instructions in experiment 2. Unfortunately, this leaves unexplained the odd pattern in effect sizes for the belief index. Why should the 10 second groups differ to a greater extent than the unspeeded groups? An alternative explanation follows from the design of experiment 1. As in Evans and Curtis-Holmes' design, the 10 second group was presented with premises in isolation, with the conclusion onsetting halfway through each trial, while for the unspeeded (and 60 second) groups the present experiment deviated from the prior design in that premises and conclusion were presented simultaneously. This was done in order to render conditions comparable to traditional belief bias preparations, allowing a more valid assessment of the potential effect of imposing a deadline with respect to previous work. It is possible, though, that the design of Evans and Curtis-Holmes is to some extent similar to a production task in that subjects were more likely to engage in premise-based (forward) reasoning as opposed to conclusion-based (backward) reasoning, in which subjects do not attempt to integrate the premise terms until after the conclusion has been read (cf. Morley et al., 2004). As mentioned in the introduction, Morley et al. (2004) have demonstrated that production tasks minimize belief bias effects relative to evaluation tasks. If the 10 second and unspeeded groups of experiment 1 can be seen as approximating forward and backward reasoning tasks, respectively, then the confound could to some extent reduce belief bias in the 10 second group of the present experiment relative to the unspeeded group, which may have also seen a correspondingly increase in the effect. This would account for both the exaggeration in differences between effect sizes of the 10 second and unspeeded groups of the two studies, as well as the marginal status of the effect of extra time in the unspeeded group by the standards of the Bonferroni correction. This is not altogether far-fetched so long as one accepts the notion of Morley et al. that the effect of conclusion-based reasoning is to bias the representation of the premises, which would be especially hard to imagine in the 10 second group, for which less than 5 seconds would be available for subjects to reconsider them.

In any case, the possibility of such confounding influences suggests a profitable direction for future work might be to examine separately the effects of response deadlines and conclusion onsets, as well as comparing the effects of the present instruction procedure with the 'standard' technique of Newstead et al. (1992).

Finally, visual inspection of the form of belief and logic ROCs suggests different models for belief-based and logic-based responding. For the 10 second group, where system 1 should have predominated in determining responses, the belief ROC was substantially above the chance line, and appeared to be more curvilinear than the logic ROCs of the 60 second and unspeeded groups. This suggests system 1 may be sensitive to gradients in believability; the best-fitting model for heuristic processing, then, might be one that assumes a continuous strength variable, such as unequal-variance detection theory. When logic-based responding predominated, as in the 60 second and unspeeded groups, the logic ROCs were substantially above the chance line, and appeared to exhibit two-piece linearity, which may suggest subjects experience some difficulty in making fine discriminations in response to the logic dimension, despite being relatively consistent in separating valid and invalid arguments. Whether this necessarily implies a threshold model assuming a small number of discrete states (e.g. Krantz, 1969) or a detection model assuming criterion variability (e.g. Mueller & Weidemann, 2008) is an open question. Clearly, an important next step in research on heuristic and analytic decision-making is the application of models assuming fundamentally different underlying processes. ROCs will be an important part of such a venture, providing both a testing ground for the assumptions of new models of reasoning, as well as helping researchers to avoid erroneous conclusions that may result from inappropriately assuming threshold statistics as measures of logical competence.

APPENDIX A

INSTRUCTIONS FOR INDUCTION AND DEDUCTION

Induction Instructions

In this experiment, we are interested in people's reasoning.

For each question, you will be given some information that you should assume to be true. This will appear ABOVE a line. Then you will be asked about a conclusion sentence BELOW the line. First, you will be asked whether the conclusion is strong or not strong. By "strong", we mean that assuming the information above the line is true, this makes the sentence below the line *plausible*. Second, you will be asked how confident you are in this judgment.

You should just answer each question as best as you can, based on the information available.

Please ask the experimenter if you have any questions.

(insert problem here)

Assuming the information above the line is true, does this make the sentence below the line *plausible*?

NOT STRONG or STRONG
(F) (J)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Press # key: 1 2 3

Deduction Instructions

In this experiment, we are interested in people's reasoning.

For each question, you will be given some information that you should assume to be true. This will appear ABOVE a line. Then you will be asked about a conclusion sentence BELOW the line. First, you will be asked whether the conclusion is valid or not valid. By "valid", we mean that assuming the information above the line is true, this *necessarily* makes the sentence below the line true. Second, you will be asked how confident you are in this judgment.

You should just answer each question as best as you can, based on the information available.

Please ask the experimenter if you have any questions.

(PROBLEM HERE)

Assuming the information above the line is true, does this *necessarily* make the sentence below the line true?

Not VALID or VALID
(F) (J)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Press # key: 1 2 3

APPENDIX B

CONCLUSION RATINGS FOR NEW CONTENT

Believable	Mean	SD	Unbelievable	Mean	SD
Some animals are not llamas	4.55	1.21	Some llamas are not animals	1.00	0.00
Some bears are not grizzlies	4.75	0.84	Some grizzlies are not bears	1.52	1.21
Some birds are not parrots	4.68	1.06	Some parrots are not birds	1.19	0.79
Some boats are not canoes	4.35	1.31	Some canoes are not boats	1.86	1.56
Some cars are not oldsmobiles	4.19	1.56	Some oldsmobiles are not cars	1.43	0.96
Some criminals are not robbers	4.61	1.05	Some robbers are not criminals	2.11	1.59
Some dances are not tangos	4.68	0.90	Some tangos are not dances	1.65	1.23
Some drinks are not beers	4.82	0.77	Some beers are not drinks	1.58	1.36
Some horses are not ponies	3.68	1.63	Some ponies are not horses	2.42	1.78
Some insects are not spiders	4.58	1.09	Some spiders are not insects	2.07	1.56
Some killers are not assassins	3.96	1.69	Some assassins are not killers	1.32	0.79
Some plants are not weeds	4.52	1.15	Some weeds are not plants	2.29	1.58
Some relatives are not uncles	4.84	0.73	Some uncles are not relatives	2.29	1.67
Some reptiles are not lizards	4.39	1.29	Some lizards are not reptiles	1.48	1.06
Some storms are not blizzards	4.86	0.76	Some blizzards are not storms	1.55	1.15
Some trees are not oaks	4.55	1.23	Some oaks are not trees	1.96	1.50
Some weapons are not cannons	4.61	1.17	Some cannons are not weapons	2.61	1.73
Some words are not verbs	4.86	0.76	Some verbs are not words	1.55	1.36
Some writers are not novelists	4.79	0.79	Some novelists are not writers	1.84	1.49

New conclusions were selected from a pool of 96 believable and unbelievable conclusions rated in a previous study; a 5-point scale was used, in which a 1 corresponded to 'Unbelievable' and a 5 corresponded to 'Believable.' One sample *t* tests indicate that the selected believable conclusions are rated as more believable than the unbelievable ones ($p<.001$), and the ratings for believable conclusions are neither more nor less variable than ratings for unbelievable ones ($p=.19$).

APPENDIX C

PROBLEM STRUCTURES USED IN EXPERIMENTS 1 AND 2

A

Set A		Set B	
Valid	Invalid	Valid	Invalid
EI2_O1	EI2_O2	OA2_O2	OE2_O2
EI3_O1	EI3_O2	AO2_O1	EO2_O1
EI4_O1	EI4_O2	OA3_O1	OE3_O1
IE4_O2	IE4_O1	AO3_O2	EO3_O2

B

No X are Y
Some Z are Y

Some Z are not X

A) Structures are identified by quantifiers used in the premises, with the first letter corresponding to the first premise and the 3rd letter corresponding to the conclusion. Following the quantifiers for the two premises will be a number corresponding to figure, and following the quantifier for the conclusion will be a number corresponding to the ordering of conclusion terms. A 1 indicates a conclusion in the Z-X direction and a 2 indicates a conclusion in the X-Z direction. B) Using this notation, the above example would be syllogism EI2_O1.

APPENDIX D

PREPARATION INSTRUCTIONS

Experiment 1 (All Subjects) and Experiment 2 (Deduction)

In the experiment, you will be asked to judge whether some conclusions are logically valid. By logically valid, we mean that the conclusion must be true, after you take account of the given information.

The given information is shown above the line, and the conclusion is shown below the line. For example,

```
All shamuses are theurgists
Some cowboys are shamuses
-----------------------------
Some cowboys are theurgists
```

Given the fact that all shamuses are theurgists, and some cowboys are shamuses, it must be true that some cowboys are theurgists. So this conclusion is valid. Why?

Here's another example.

```
All carolingians are paladins
All rulers are carolingians
-----------------------------
All rulers are paladins
```

Given that all carolingians are paladins, and all rulers are carolingians, it must be true that all rulers are paladins. So this conclusion is valid. Why?

Now consider this example.

```
All karrozzins are hammerkops
No karrozzins are sculptors
---------------------------------
All sculptors are hammerkops
```

Given the fact that all karrozzins are hammerkops, and no karrozins are sculptors, you can't conclude that all sculptors must be hammerkops. So, this conclusion is not valid. Why?

In this experiment, it is very important that you only say that a conclusion is valid when it must be true given the information above the line. If the conclusion is not necessarily true, then say not valid.

Please ask the experimenter if you have any questions.

Preparation Instructions: Experiment 2 (Induction)

In the experiment, you will be asked to judge whether some conclusions are strong. By strong, we mean that the conclusion is plausible, after you take account of the given information.

The given information is shown above the line, and the conclusion is shown below the line. For example,

```
All shamuses are theurgists
Some cowboys are shamuses
-----------------------------
Some cowboys are theurgists
```

Given the fact that all shamuses are theurgists, and some cowboys are shamuses, it is plausible that some cowboys are theurgists. So this conclusion is strong. Why?

Here's another example.

```
All carolingians are paladins
All rulers are carolingians
-------------------------------
All rulers are paladins
```

Given that all carolingians are paladins, and all rulers are carolingians, it is plausible that all rulers are paladins. So this conclusion is strong. Why?

Now consider this example.

```
All karrozzins are hammerkops
No karrozzins are sculptors
---------------------------------
All sculptors are hammerkops
```

This conclusion is not strong. Given the fact that all karrozzins are hammerkops, and no karrozzins are sculptors, it's not plausible that all sculptors are hammerkops. Why?

In this experiment, it is very important that you only say that a conclusion is strong when it is plausible given the information above the line. If the conclusion is not likely, then say not strong.

Please ask the experimenter if you have any questions

APPENDIX E

DEADLINE PRACTICE INSTRUCTIONS

In this experiment, you will have (**insert 10 seconds, 16 seconds, or 1 minute**) to respond 'Valid' or 'Invalid'. A timer will indicate how much time is left before a response must be made. If you do not respond in time, you will be advanced automatically to the next trial.

After you make a response, you will be asked how confident you are that the response was correct. Your confidence rating will not be timed, however, and you should use this time wisely to accurately indicate your rating.

APPENDIX F

PRACTICE PROBLEMS FOR EXPERIMENT 2

Deduction

Welcome to the experiment! In this study, we are interested in people's reasoning. You will be asked to respond to several short logic problems; some of them will be rather easy and some may be a bit more complex. In any case, just try to do the best that you can. Below is an example of what you will see in the experiment; to be sure you understand the task before engaging in the experiment, please try the practice problems below and be sure to ask the experimenter if you have any questions.

For each question, you will be given some information that you should assume to be true. This will appear ABOVE a line. Then you will be asked about a conclusion sentence BELOW the line. First, you will be asked whether the conclusion is valid or not valid. By “valid”, we mean that assuming the information above the line is true, this *necessarily* makes the sentence below the line true. Second, you will be asked how confident you are in this judgment.

You should just answer each question as best as you can, based on the information available.

Please ask the experimenter if you have any questions.

Example Problem 1

No invectives are critiques
Some invectives are vituperations

Some vituperations are not critiques

Assuming the information above the line is true, does this *necessarily* make the sentence below the line true?

NOT VALID or VALID

(Circle one)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Circle one: 1 2 3

Example Problem 2

All chameleons are squamates
Some coxcombs are squamates

Some chameleons are coxcombs

Assuming the information above the line is true, does this *necessarily* make the sentence below the line true?

NOT VALID or VALID

(Circle one)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Circle one: 1 2 3

Induction

Welcome to the experiment! In this study, we are interested in people's reasoning. You will be asked to respond to several short logic problems; some of them will be rather easy and some may be a bit more complex. In any case, just try to do the best that you can. Below is an example of what you will see in the experiment; to be sure you understand the task before engaging in the experiment, please try the practice problems below and be sure to ask the experimenter if you have any questions.

For each question, you will be given some information that you should assume to be true. This will appear ABOVE a line. Then you will be asked about a conclusion sentence BELOW the line. First, you will be asked whether the conclusion is strong or not strong. By "strong", we mean that assuming the information above the line is true, this makes the sentence below the line *plausible*. Second, you will be asked how confident you are in this judgment.

You should just answer each question as best as you can, based on the information available.

Please ask the experimenter if you have any questions.

Example Problem 1

No invectives are critiques
Some invectives are vituperations

Some vituperations are not critiques

Assuming the information above the line is true, does this make the sentence below the line *plausible*?

NOT STRONG or STRONG

(Circle one)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Circle one: 1 2 3

Example Problem 2

All chameleons are squamates
Some coxcombs are squamates

Some chameleons are coxcombs

Assuming the information above the line is true, does this make the sentence below the line *plausible*?

NOT STRONG or STRONG

(Circle one)

How confident are you in this judgment?

1=not at all confident, 2=moderately confident, 3=very confident

Circle one: 1 2 3

REFERENCES

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology, 53*(1), 77-86.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*(2), 137-181.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*(2), 81-99.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 929-945.

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81*(2), 351-354.

Beller, S., & Spada, H. (2003). The logic of content effects in prepositional reasoning: The case of conditional reasoning with a point of view. *Thinking & Reasoning, 9*(4), 335-378.

Chater, N., & Oaksford, M. (2001). Human rationality and the psychology of reasoning: Where do we go from here? *British Journal of Psychology, 92*(1), 193-216.

Chen, Z., & Daehler, M. (2000). External and internal instantiation of abstract information facilitates transfer in insight problem solving. *Contemporary Educational Psychology, 25*(4), 423-449.

Cherubini, P., Garnham, A., Oakhill, J., & Morley, E. (1998). Can any ostrich fly? some new data on belief bias in syllogistic reasoning. *Cognition, 69*(2), 179-218.

Copeland, D., & Radvansky, G. (2004, November). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 57A*(8), 1437-1457.

Copi, I., & Cohen, C. (1994). *Introduction to logic.* Upper Saddle River, NJ, US: Prentice Hall.

Dickstein, L. S. (1975). Effects of instructions and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 1*(4), 376-384.

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory & Cognition, 6*(1), 76-83.

Dickstein, L. S. (1981). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18*(5), 229-232.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*(1), 91-101.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*, (58), ii.

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*(3), 295-306.

Evans, J. S. B. T., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology, 6*(3), 263-285.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13(3)*, 378–395.

Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. New York, NY, US: Psychology Press.

Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*(4), 382-389.

Evans, J. S. B. T., Handley, S. J., & Harper, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 54*(3), 935-958.

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Feeney, A. (2007). Individual differences, dual processes, and induction. In A. Feeney, & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches.* (pp. 302-327). New York, NY, US: Cambridge University Press.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54*(1), 1-71.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105*(3), 331-351.

Gilinsky, A. S., & Judd, B. B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging, 9*(3), 356-371.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 500-513.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87*(1), B11-b22.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.

Grice, H. P. (2002). Logic and conversation. In D. J. Levitin (Ed.), Foundations of cognitive psychology: Core readings (pp. 719-732). Cambridge, MA: MIT Press. (Reprinted from Syntax and semantics 3: Speech acts, pp. 26-40, by P. Cole & J. L. Morgan, Eds., 1975, New York: Academic Press).

Hawkins, J., Pea, R. D., Glick, J., & Scribner, S. (1984). 'Merds that laugh don't like mushrooms': Evidence for deductive reasoning by preschoolers. *Developmental Psychology, 20*(4), 584-594.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1210-1230.

Heit, E. (2007). What is induction and why study it? In A. Feeney, & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches.* (pp. 1-24). New York, NY, US: Cambridge University Press.

Heit, E., & Rotello, C.M. (2005). Are there two kinds of reasoning? *In Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language, 54*, 425-443.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press. Cambridge, Eng.: Cambridge University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*(1), 1-61.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10*(1), 64-99.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 701-722.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*(4), 852-884.

Krantz, D. (1969). Threshold theories of signal detection. *Psychological Review*, *76*(3), 308-324.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17*(1), 11-17.

Markovits, H., & Schroyens, W. (2007). A curious belief-bias effect: Reasoning with false premises and inhibition of real-life information. *Experimental Psychology, 54*(1), 38-43.

Markovits, H., & Vachon, R. (1989). Reasoning with contrary-to-fact propositions. *Journal of Experimental Child Psychology, 47*(3), 398-412.

Markovits, H., & Vachon, R. (1990). Conditional reasoning, representation, and level of abstraction. *Developmental Psychology, 26*(6), 942-951.

Morley, N. J., Evans, J. S. B. T., & Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 57*(4), 666-692.

Mueller, S., & Weidemann, C. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465-494.

Newstead, S. E., Pollard, P., & Evans, J. S. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*(3), 257-284.

Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31*(2), 117-140.

Polk, T., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102*(3), 533-566.

Quayle, J., & Ball, L. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 53A*(4), 1202-1223.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763-785.

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518-535.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12*(2), 129-134.

Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief bias effect in formal reasoning: The influence of knowledge on logic. *Memory & Cognition, 8*(6), 584-592.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior, 14*, 180-195.

Roberts, M., Newstead, S., & Griggs, R. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173-204.

Roberts, M. J., & Sykes, E. D. A. (2003). Belief bias and relational reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 56*(1), 131-154.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review, 111*(3), 588-616.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389-401.

Sells, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology (Columbia University),* (200), 72-72.

Schmidt, J., & Thompson, V. (2008). 'At least one' problem with 'some' formal reasoning paradigms. *Memory & Cognition*, *36*(1), 217-229.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491-517.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*(3), 619-632.

Sidman, M. (1960). *Tactics of scientific research*. Oxford, England: Basic Books.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*(2), 231-280.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*(1), 1-33.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-726.

Thompson, V. (1996). Reasoning from false premises: The role of soundness in making logical deductions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*(3), 315-319.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review*, *10*(1), 184-189.

Torrens, D., Thompson, V., & Cramer, K. (1999). Individual differences and the belief bias effect: Mental models, logical necessity, and abstract reasoning. *Thinking & Reasoning*, *5*(1), 1-28.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*(4), 451-460.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341-1354.